Master's thesis
Master's Programme in Data Science

Probabilistic Predictive Elicitation

Georgi Agiashvili

June 13, 2021

Supervisor(s): Assistant Professor Arto Klami

Examiner(s): Assistant Professor Arto Klami
Dr. Marcelo Hartmann

University of Helsinki
Faculty of Science

P. O. Box 68 (Pietari Kalmin katu 5)
00014 University of Helsinki

HELSINGIN YLIOPISTO — HELSINGFORS UNIVERSITET — UNIVERSITY OF HELSINKI

Tiedekunta — Fakultet — Faculty	Koulutusohjelma — Utbildningsprogram — Degree programme
Faculty of Science	Master's Programme in Data Science

Tekijä — Författare — Author

Georgi Agiashvili

Työn nimi — Arbetets titel — Title

Probabilistic Predictive Elicitation

Työn laji — Arbetets art — Level	Aika — Datum — Month and year	Sivumäärä — Sidantal — Number of pages
Master's thesis	June 13, 2021	82

Tiivistelmä — Referat — Abstract

Unlike the traditional machine learning approaches that rely solely on data, Bayesian machine learning models can utilize prior knowledge on the data generating process, for instance in form of information about plausible outcomes. More importantly, Bayesian machine learning models use the prior information as the base knowledge, on top of which the learning from observations is built on. The process of forming the prior distribution based on subjective probabilities is called prior elicitation, and that is the focus of this thesis.

Although previous research has produced methods for prior elicitation, there has not been a general-purpose solution. Particularly, the methods introduced previously have focused on specific models. This has limited the applicability of prior elicitation, and in some cases, required the expert to have a deep understanding of different aspects of the Bayesian modelling. Additionally, the more general predictive elicitation methods in previous research have not accounted for the uncertainty regarding experts' judgements. This is important, since even the most accurate elicitation methods cannot remove all imprecision in expert judgements. Because of these reasons, prior elicitation has remained somewhat underrated and underused in the modern Bayesian workflow.

This thesis provides a theoretical basis and validation of a novel prior elicitation method, which was first introduced by Hartmann et al. [37]. Particularly, this principled statistical framework called probabilistic predictive elicitation 1) makes prior elicitation independent on the specific structure of the probabilistic model, 2) handles complex models with many parameters and potentially multivariate priors, 3) fully accounts for uncertainty in experts' probabilistic judgements on the data, and 4) provides a formal quality measure indicating if the chosen predictive model is able to reproduce experts' probabilistic judgements.

We extend the published work [37] in multiple ways. First, we provide more thorough literature reviews on different prior elicitation approaches as well as on methods for the expert elicitation. Second, we continue the discussion about technicalities, implementation and applications of the proposed methodology. Third, we report two unpublished experiments using the proposed methodology. In addition, we discuss the methodology in the context of the modern Bayesian workflow.

ACM Computing Classification System (CCS):
Computing methodologies → Modeling and simulation → Model development and analysis → Uncertainty quantification

Avainsanat — Nyckelord — Keywords

prior elicitation, predictive elicitation, Bayesian modelling, prior predictive distribution

Säilytyspaikka — Förvaringsställe — Where deposited

Muita tietoja — Övriga uppgifter — Additional information

Contents

1. Introduction

The vast amount of data collected nowadays has lead the machine learning research to focus on how to utilize this computer-stored data in training models for various purposes, such as medical treatment recommendation, targeted marketing and risk management. However, in many real world problems the practitioners have found that not all important data is available, nor is all of the available data reliable. One solution for addressing these issues is the probabilistic, or Bayesian, approach to machine learning. In Bayesian machine learning, all uncertainty regarding the data is described in form of probabilities that explicitly encode uncertainty.

A distinctive property of Bayesian models is that they include information about the modelled topic prior to receiving any observations. The prior information can be based on previous data or, ideally, on expert knowledge of the subject matter. In other words, such machine learning models can incorporate a base knowledge before processing data. This is unlike the currently trending branch of machine learning that utilizes primarily the possibilities of big data, for instance in form of deep neural networks trained to recognize objects from images. In addition to accounting for uncertainty of the data, probabilistic models can be validated by using the prior knowledge.

Recently, the research community has focused on identifying guidelines for a good workflow for building Bayesian models [10, 79]. This is due to the increasing recognition of the advantages of robust Bayesian analysis [32], as well as to the rising computing power and the progress made in the development of probabilistic programming languages such as `Stan` [12], `WinBUGS` [55] and `JAGS` [76]. The modern Bayesian workflow aims to provide tools for answering the following four questions [10]:

- Is our model consistent with our domain expertise?
- Will our computational tools be sufficient to accurately fit our posteriors?
- Will our inferences provide enough information to answer our questions?
- Is our model rich enough to capture the relevant structure of the true data generating process?

It is worth noting, that only the last question requires actual observations to answer. For example, in principled workflow described by Betancourt [10], 11 out of 14 concrete steps are done in pre-data phase.

This thesis focuses on principled deployment of domain knowledge for Bayesian analysis before obtaining observations. From the perspective of the Bayesian workflow, the point of interest here is the consistency of the model with the domain expertise. This is done through *predictive elicitation* [1, 30, 46, 87], which translates the expert's implicit knowledge, and more precisely the predictions of the outcomes of the generative process, into explicit quantities that are modelled as the prior knowledge. This complements the current guidelines of the Bayesian workflow, where visual and qualitative prior predictive checks are recommended [10, 25], by providing means for a systematic model validation. Particularly, the methodology introduced by Hartmann et al. [37] and extended further in this thesis allows incorporating the prior knowledge into a probabilistic model while accounting for the uncertainty of the expert's knowledge.

While using elicitation in modern Bayesian workflow is challenged for being expensive and imprecise [10], the methodology introduced in this thesis requires from the expert little more than what is needed for predictive checking. The methodology, called *probabilistic predictive elicitation*, has two key features that make it particularly useful for the Bayesian workflow. First, it accounts for the noise in expert knowledge. The quantification of the expert knowledge, i.e. *expert elicitation*, is never perfect [70] even though it has been widely studied [69]. Second, the methodology is model-independent unlike the methods introduced in previous literature. This feature provides flexibility in two forms. The expert does not need to know the modelling aspects while providing probabilistic judgements of the outcomes of generative process. In addition, since the expert provides judgements independently from the statistical model, our methodology can be used to compare different models in the pre-data phase of the Bayesian workflow.

The main contributions of this thesis, and the research paper *Flexible Prior Elicitation via the Prior Predictive Distribution* [37] published in the Conference on Uncertainty in Artificial Intelligence in 2020, are the theoretical basis as well as technical tools for applying the probabilistic predictive elicitation. Particularly, this novel statistical framework 1) makes prior elicitation independent on the specific structure of the probabilistic model, 2) handles complex models with many parameters and potentially multivariate priors, 3) fully accounts for uncertainty in experts' probabilistic judgements on the data, and 4) provides a formal quality measure indicating if the chosen predictive model is able to reproduce experts' probabilistic judgements.

This thesis goes beyond our research paper [37] in multiple ways. Here, we provide more thorough literature reviews on different prior probability elicitation approaches,

as well as on methods for (human) expert elicitation. We also provide supplementary graphs and discussion of the real experiment used in the original paper (Section 6.2, and apply the new methodology in two additional experiments (Sections 6.1 and 6.3). The first of the new experiments was included in the first draft of the original paper, while the second provides an original study and shows new opportunities for the Bayesian workflow. While the description of the methodology in this thesis generally follows the original paper, we extend that with more discussion and explanation about the technicalities.

Chapter 2 provides introduction to Bayesian inference and prior elicitation. Chapter 3 discusses expert elicitation in more general terms and provides reasoning why the uncertainty of experts judgements should be considered in prior elicitation. Chapter 4 introduces the new methodology for prior elicitation and discusses its properties, while Chapter 5 provides mathematical tools to handle the computation of elicitation process. Chapter 6 presents applications of the new method, and Chapter 7 concludes.

2. Bayesian Inference and Prior Elicitation

In Bayesian inference, the *Bayes' theorem* is applied to update model parameters as observations are made. According to the Bayesian paradigm, only the observations are known while all the other elements of the Bayesian model, such as parameters and model structure, are uncertain. The underlying uncertainties are modelled as probabilities, which are interpreted as the degree of belief, i.e. how likely a certain outcome would occur[†]. Throughout the thesis this Bayesian interpretation of probabilities is implied when discussing probabilistic inference or probabilistic approach.

The defining characteristic of Bayesian modelling is that the model parameters are treated as uncertain values. Therefore, they are assigned with prior probability distributions before making any observations. This is unlike in the frequentist statistics, where only the observed data describes the model parameters. Hence, the prior distribution implies the prior knowledge about the (inference) problem. There are different approaches to forming prior distributions which will be discussed shortly. However, the key concept regarding this thesis is *prior elicitation*, which means the process of forming the prior distribution based on subjective probabilities of a single or multiple experts.

In this chapter, we will first give a general notation to the basic concepts in Bayesian inference. Then, we discuss the properties and interpretations of the prior distribution. Finally, we go through the previous work on prior elicitation.

2.1 Basics of the Bayesian inference

Suppose that the observations y follow a normal distribution, meaning that the *observation model* is $\mathcal{N}(y|\mu, \sigma^2)$. While in frequentist statistics we assume that model parameters μ and σ are determined only by the observations, in *Bayesian model* the

[†]As opposed to the frequentist approach where the interpretation of probability is the relative frequency or how often a certain outcome would occur.

assumption is that there is uncertainty regarding those parameters. Thus, μ and σ follow some probability distribution, and the purpose of the Bayesian inference is to find that distribution instead of some specific parameters. For example, before getting any observations we could assume that $\mu \sim \mathcal{N}(m, s^2)$ and $\sigma \sim \text{Gamma}(a, b)$. Then, we are interested in obtaining the distribution $P(\mu, \sigma|y)$ for model parameters that are conditioned on the observations. Also, we could further introduce prior and (conditioned) posterior uncertainty in the form of probabilities to the *hyperparameters* m, s, a and b. The distribution we assign to a model parameter before considering observations is the *prior probability distribution*, whereas the inferred distribution that combines both, the prior expectations and observations, is the *posterior probability distribution*.

More formally, given the observations y and model parameters $\boldsymbol{\theta}$, a Bayesian model consists of an observation model (or *likelihood*) $\pi(y|\,\boldsymbol{\theta})$, prior probability distribution $\pi(\boldsymbol{\theta})$, posterior probability distribution $\pi(\boldsymbol{\theta}\,|y)$, and *marginal likelihood* $\pi(y)$, which is also known as a *normalizing constant*. Conditioning these components on a specific Bayesian model $\mathcal{M}$, the Bayesian inference is described by the Bayes' theorem

$$\pi(\boldsymbol{\theta}\,|y, \mathcal{M}) = \frac{\pi(y|\,\boldsymbol{\theta}, \mathcal{M})\pi(\boldsymbol{\theta}\,|\,\mathcal{M})}{\pi(y|\,\mathcal{M})}. \tag{2.1}$$

The marginal likelihood is simply the likelihood marginalized over prior

$$\pi(y|\,\mathcal{M}) = \int \pi(y|\,\boldsymbol{\theta}, \mathcal{M})\pi(\boldsymbol{\theta}\,|\,\mathcal{M})\mathrm{d}\boldsymbol{\theta}, \tag{2.2}$$

and it can be opened as a sequence of predictive distributions using the chain rule

$$\pi(y|\,\mathcal{M}) = \pi(y_1|\,\mathcal{M}) \prod_{t=2}^{T} \pi(y_t|y_{1,\ldots,t-1}, \mathcal{M}). \tag{2.3}$$

Here y_t is the data collected at a time step t and

$$\pi(y_1|\,\mathcal{M}) = \int \pi(y_1|\,\boldsymbol{\theta}, \mathcal{M})\pi(\boldsymbol{\theta}\,|\,\mathcal{M})\mathrm{d}\boldsymbol{\theta} \tag{2.4}$$

is the prior predictive distribution for realization of y. On the other hand, the posterior predictive distribution is conditioned on the previous observations and it is given by

$$\pi(y_t|y_{1,\ldots,t-1}, \mathcal{M}) = \int \pi(y_t|\,\boldsymbol{\theta}, \mathcal{M})\pi(\boldsymbol{\theta}\,|y_{1,\ldots,t-1}, \mathcal{M})\mathrm{d}\boldsymbol{\theta}\,. \tag{2.5}$$

The predictive distributions are used to describe the data generating process that is being inferred. Following the eq. (2.3), the marginal likelihood equals the prior predictive distribution when no observations are acquired. Thus, it is the *a priori* description of the data generating process.

In modern Bayesian workflow, predictive distributions are important part of prior and posterior model checking [32, 25, 79]. Here, the domain knowledge is utilized to

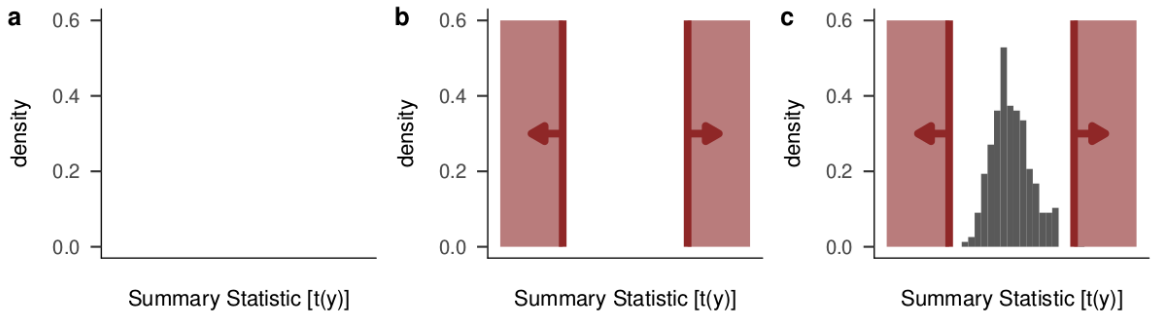


Figure 2.1: Example of prior predictive checking by Schad, Betancourt and Vasishth [79]: **a)** In a first step, define a summary statistic that one wants to investigate. **b)** Second, define extremity thresholds (shaded areas), for which one does not expect a lot of prior data. **c)** Third, simulate prior model predictions for the data (histogram) and compare them with the extreme values (shaded areas). In this example, the modeller would likely accept the model as reasonable, since the simulated data stays within the thresholds.

validate the plausibility of the model (prior) and the adequacy of the fit of the model (posterior). Simulated hypothetical data from the predictive distributions is often much easier for a domain expert to assess compared to prior and posterior distributions of the model parameters. An example of prior predictive checking is provided in Figure 2.1. Furthermore, in model comparison and selection, the predictive methods have become important in the Bayesian context [85, 75].

2.2 Prior distribution

As mentioned above, one defining characteristic of Bayesian modelling is that there are a priori assumptions of the generative model that is being inferred. More precisely, these assumptions are introduced through the prior distribution. Traditionally, Bayesian statisticians have been divided as *objectivist* and *subjectivist* Bayesianists. Although such division is argued to be misleading [33], it has set ground to how prior distributions have been formed. Subjectivists have argued that the prior distribution should express specific knowledge or subjective views about the prior before observing new information. On the other hand, objectivists have attempted to produce general-purpose methods for prior formation, which emphasizes the role of observations and the observation model in the analysis.

According to the review by Kass and Wasserman [50], priors that are formed based on the objectivist approach were referred as "noninformative", but following a more fundamental reasoning they are more suitably called *reference priors*. Correspondingly, priors formed based on the subjectivist approach were called "informative", but *subject-specific priors* offer a more descriptive label. Based on Gelman [31], *informativeness*

of the prior can be understood as a merely technical term describing how evenly the probability mass is distributed across the prior space, where more informative distribution means more concentrated probability mass and vice versa. According to this definition, subject-specific priors are at least to some extent informative, whereas the informativeness of reference priors can vary significantly depending on the approach by which they are formed.

Arguably, subject-specific priors may also be called "informative" or "generative". First, Gelman [34] defines *weakly informative priors* to contain some but not all of the subject-specific knowledge. This implies that informativeness would relate to subject-specific knowledge. Second, *generative priors* are by strict definition [34] formed so that the resulting prior predictive distribution has desirable properties. However, priors that contain subject-specific information can also be formed without assessing them in relation with the predictive distribution, as we will discuss in Section 2.3.1. Thus, calling priors that are based on subject-specific knowledge as "generative" is not always accurate, whereas calling these priors "informative" can be misleading. Since the traditionally used term subjective prior is also problematic [33], we proceed in this thesis with the term "subject-specific".

Reference priors are generally constructed by formal rules to express *ignorance*, which means that the analyst does not take a stance on the possible hypotheses. Calling formally constructed priors "noninformative", which is based on the subjectivist-objectivist dispute, is debatable because different methods may produce different inference results [50]. Kass and Wasserman [50] call the formally constructed priors as "reference priors", since they are used as the default option when choosing the prior distributions. The simplest rule for determining reference prior is the principle of indifference, in which equal probabilities are assigned to all possibilities. Kass and Wasserman [50] review a number of priors constructed by formal rules, and conclude that they are often *improper*, i.e. their sum or integral is not finite, and they depend on the experimental design, which may lead to unsatisfactory results. Nevertheless, a large number of observations will usually decrease the impact of prior as shown in Figure 2.2, which has made the use of diffuse priors widely accepted.

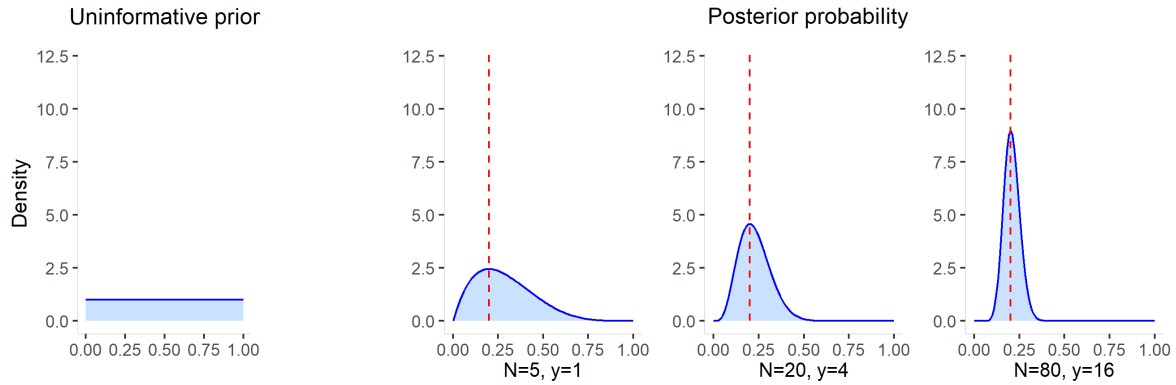
Subject-specific priors are often informative, since the knowledge used to assign probabilities can be presumably used to separate more likely values from the less likely ones. Of course, if there is no a priori knowledge, then subject-specific prior could be diffuse and thus less or even noninformative. Even when enough knowledge to form informative priors is available, Gelman [34, 31] suggests the use of weakly informative priors which merely express common sense limitations for the generative model. Suppose a model that has an outdoor temperature parameter that is being estimated. In most cases it would make little sense to allow it go too much above or below the histor-

ical extreme measurement points. Thus, a weakly informative prior would emphasize possible values from a predefined range that is based on the measurement history. However, it would still assign some probability on the values outside that range. Thus, ideally the weakly informative priors help avoiding overfitting (which could happen if the priors were too narrow) while they still include previous knowledge [34].

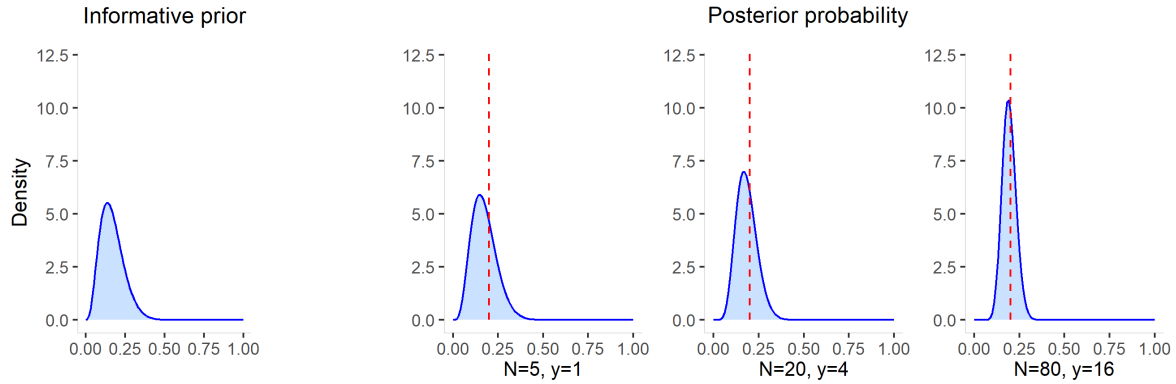
Sarma and Kay [78] surveyed practitioners of Bayesian statistics from various fields to find how they form prior distributions. The researchers chose a predefined model and allowed the participants to choose priors and their hyperparameters from normal and Student's t distributions with predefined ranges for hyperparameter values. Participants had different philosophical reasons for their choices of priors. While some leaned towards skeptical zero-centered priors that require strong evidence to conclude that the parameters have large effect, others tended towards uninformative priors with large scale parameters to account for participants' lack of previous knowledge. In addition, the researchers found that the practitioners often used the Student's t distribution as a hedge to the possibility of being wrong. On the other hand, normal distribution was used when practitioners did not express the need for such hedge, for example when setting less informative priors.

In the survey by Sarma and Kay [78], many practitioners wanted to use weakly informative priors as recommended by Gelman [34, 31]. However, the researchers found that the interpretation of such prior was inconsistent across the participants. For example, some of the participants who identified using weakly informative priors chose the smallest available value for the scale parameter of a distribution while some chose the largest value. Moreover, some of these participants chose priors that generated theoretically impossible values in predictive distribution while others avoided this potential issue. The researchers concluded that the absence of concrete guidelines for forming weakly informative priors means that such priors depend on analyst's taste and experience.

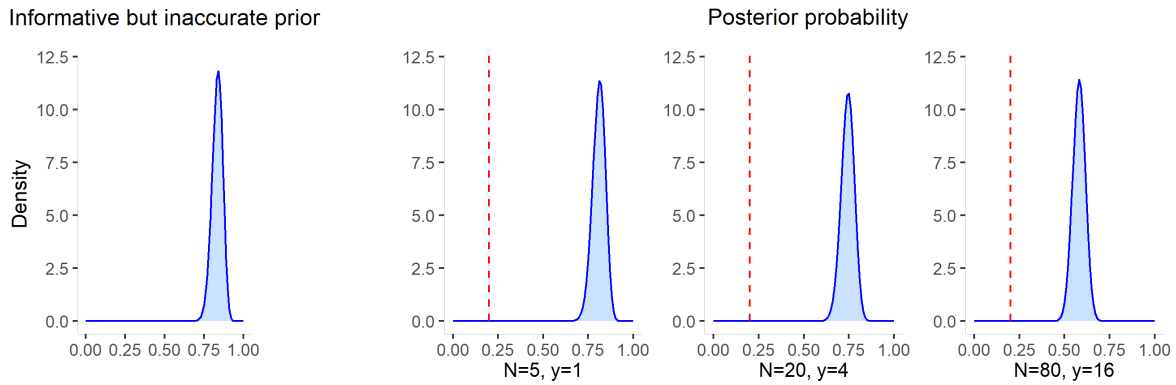
Using informative priors has several practical advantages over less informative priors. First, as shown in Figure 2.2 the informative prior can complement small sample size assuming the prior is appropriate. Second, informative prior can help with convergence in computation [33]. Third, informative priors can produce a more sensible *Bayes factor*, which is the ratio of marginal likelihoods of two Bayesian models primarily used for model comparison. Particularly, this happens by avoiding the *Bartlett's paradox* [6, 49], where noninformative or improper prior forces to favor one model over the other based on the Bayes factor. However, when using a specific informative prior a great care should be taken on the accuracy of the prior since an inaccurate prior may hinder the inference [21]. This is also shown in the Figure 2.2. Furthermore, too informative priors may overfit the model decreasing its predictive performance [34].



(a) Posterior probability densities with an uninformative prior $\pi(p) \sim \text{Beta}(1, 1)$.



(b) Posterior probability densities with an informative prior $\pi(p) \sim \text{Beta}(4, 20)$.



(c) Posterior probability densities with an inaccurate informative prior $\pi(p) \sim \text{Beta}(100, 20)$.

Figure 2.2: A simple illustration how the choice of prior and the number of observations affect the posterior distribution. Here, we have an observation model $\pi(y|p) = \text{Binom}(y|N, p)$ with differently parameterized beta-distributed priors. In this example, the true $p = 0.2$, shown as the red dashed line, and observations are assumed accordingly. **a) and b)** Use of the informative prior produces a sharp posterior distribution with fewer observations than the uninformative prior, but as the number of observations grows, the impact of prior decreases as illustrated from left to right. **c)** Inaccurate prior hinders the inference, however, as the number of observations grows, the posterior distribution moves towards the correct distribution. However, for a small number of observations N there are critical issues with the inaccurate model: not only is the mode of the distribution poor, but the prior probability of the true p is close to zero, and the narrow peak gives a false impression of reliability.

Regardless of the approach chosen to form a prior, ideally, the prior should be designed independently of observations [9, 34, 10]. If the observations would be used while setting the prior and then again when constructing the posterior distribution, the observed data would over-influence the model, making the model prone to overfitting. Thus, the prior should at most reflect the a priori knowledge about the modelled issue. In practice, this is easier to achieve with reference priors due to their formal nature. In that case, posterior predictive checking for model validation is encouraged in practical inference [34], which can be seen to violate the independence between the prior and observations. Construction of a subject-specific prior independently of observations is done by quantifying the information from previous knowledge. This means either the use of previously inferred results, or the elicitation of an expert's views.

2.3 Prior elicitation

Prior elicitation is a process where a subject-specific prior for the model is elicited from an expert with a domain knowledge on the modelled topic. Note that the expert is not restricted to be an individual with vast domain-knowledge, instead, here the term *expert* refers to any individual, a group of individuals or a system that is capable of providing some information for elicitation. Prior elicitation can be divided into two distinct approaches, structural and predictive elicitation [46, 47]. In structural elicitation the expert is asked directly about the prior distribution, after which the prior is formed based on the expert's judgements. This requires the expert to understand the parameterization of the observation model, and thus it is *model-dependent*. On the contrary, in predictive elicitation the expert is asked about prior predictive distribution, and based on the expert-given probabilities the prior distribution is computed. This approach is in principle *model-independent*, although most of the existing literature about predictive elicitation focuses on some distinct set of models. The method introduced in this thesis (Chapter 4) is based on the predictive approach for prior elicitation with an emphasis on the model-independence it provides.

Kadane and Wolfson [45] provide desiderata for elicitation, which is shown in Table 2.1. Although the desiderata support mainly the predictive approach for elicitation, structural elicitation can be argued to fit the desiderata in some applications, particularly in econometrics. This is reasonable when the parameters, for which prior distributions are elicited for, are interpretable and of expert's knowledge. However, in most cases of the structural approach the desiderata is not met, since it requires the expert to be able to think parametrically. In other words, the expert would have to understand aspects of modelling in addition to their domain expertise. Therefore, the prior predictive distribution qualifies more often as a probability distribution for

Desiderata for elicitation:

(a) expert opinion is the most worthwhile to elicit;

(b) experts should be asked to assess only observable quantities, conditioning only on covariates (which are also observable) or other observable quantities;

(c) experts should not be asked to estimate moments of a distribution (except possibly the first moment); they should be asked to assess quantiles or probabilities of the predictive distribution;

(d) frequent feedback should be given to the expert during the elicitation process;

(e) experts should be asked to give assessments both unconditionally and conditionally on hypothetical observed data

Table 2.1: Kadane and Wolfson's [45] desiderata for elicitation.

observable quantities than the prior distribution of a model parameter.

2.3.1 Structural elicitation

In structural elicitation the expert is asked about possible values for parameters of the observation model. In some fields, such as economics, where linear models are a golden standard, thinking parametrically in terms of coefficients is plausible. Nevertheless, even in economics the suitability of a linear model to express data generating process becomes questionable. Moreover, as explained in [46], even if the number of variables is restricted so that they are independent and the model indeed is linear, the interpretation that the coefficients capture only the effect of the chosen variables is questionable, thus casting a doubt on the suitability of a coefficient to be the subject of elicitation. Furthermore, when adding a control variable, it can have a dependency with some of the other variables. Therefore, the *ceteris paribus* line of thinking, where change in one coefficient is allowed while keeping all others still, may lead to very peculiar results in elicitation.

Structural elicitation has been tightly related to fitting elicited information into a predefined parametric distribution. In early studies this was attempted through asking the expert about moments of a distribution. That was quickly found problematic due to expertise required and cognitive biases, which are discussed more in Section 3.1. Thus, the practice has shifted towards asking the expert about probability regions for the quantity of interest, which in structural elicitation is a parameter of the observation model. Also, to tackle the question of suitability of *ceteris paribus* approach in

elicitation, methods to include dependency in elicitation have been proposed [19].

A more recent literature in structural elicitation has focused on eliciting nonparametric distribution for the prior by fitting Gaussian processes to the elicited probabilities [66, 36, 58]. This allows flexibility for the elicitation in the sense that expert's information can be more accurately modelled without being a subject of restrictions of parametric probability distributions. Yet, nonparametric priors can be hard to use with the existing tools for Bayesian modelling such as `Stan` [12] or `brms` [11], because they assume the priors correspond to some standard distributions implemented in their interface. Moreover, the interpretation of nonparametric priors is not always easy since they are stochastic processes rather than fixed-dimensional probability distributions [88].

In practice, structural elicitation has been useful in applications that require hierarchical modelling [45, 67]. Indeed, in hierarchical models the priors often have an observable interpretation, making the structural elicitation justified. Moreover, in applications where there is no data available at all for observable priors, such as risk analysis, structural elicitation has proven valuable [64].

2.3.2 Predictive elicitation

In predictive elicitation the expert is asked about possible values of data generated by the process of interest. This data is presumably observable, thus fulfilling the fundamental requirement of successful elicitation. Predictive elicitation is inherently model-independent and requires the expert only to possess domain knowledge. This is to say, the expert is not expected to have statistical or modelling experience nor do they even need to know the model that is being used.

Although being in theory model-independent, in practice, predictive elicitation has been difficult to formalize and it has usually been developed for specific purposes, particularly due to the computational complexity. Early work in predictive elicitation focused in finding priors for Bernoulli [86] and normal linear models [28, 48]. Later, methods for predictive elicitation in generalized linear models (GLMs) were proposed [7, 27]. An example of a predictive elicitation tool for GLMs is the `Elicitator` [42] illustrated in Figure 2.3. Other models for which predictive elicitation methods are introduced include multivariate normal models [2] and generalized extreme value distributions [26].

Percy [72, 73, 74] suggested a predictive prior elicitation methodology for families of conjugate distributions. There, a conjugate prior distribution is chosen based on the sampling distribution i.e. a likelihood that must be a parametric probability distribution. The expert is asked to provide probabilities for regions in the prior pre-

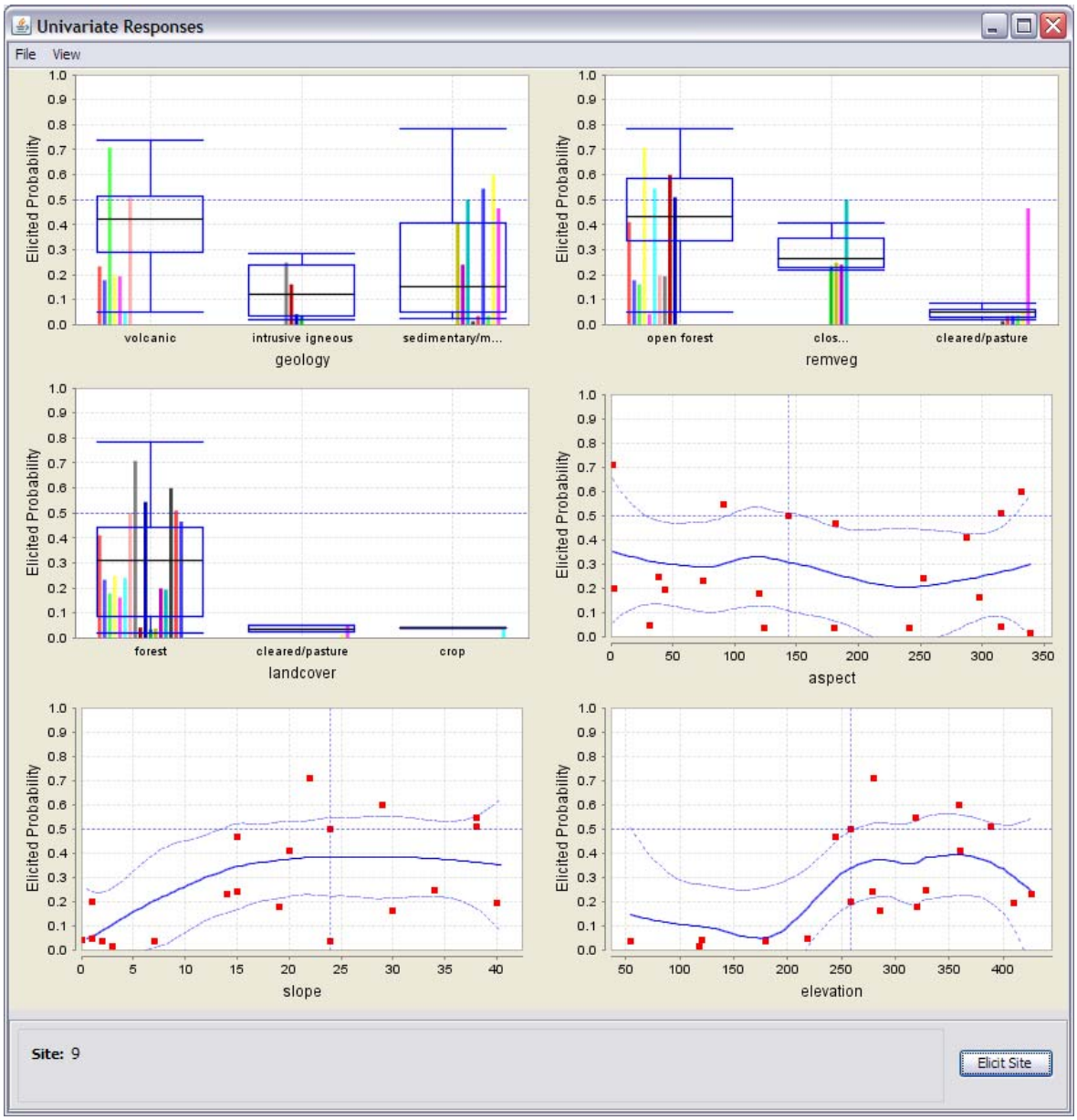


Figure 2.3: `Elicitator` [42] provides an extended implementation of the method proposed by Bedrick, Christensen and Johnson [7] for predictive elicitation of GLMs. The depicted *response viewer* shows visualization of a logistic regression. Each graph represents a different covariate, where y-axis depicts predicted probabilities of success with respect to the model's covariate values in x-axis. For categorical covariates, the expert's predictive probabilities are shown as bar charts, while box and whisker plots show the summarized prior predictions. For continuous covariates, the expert's predictive probabilities are represented by dots positioned at the covariate value, while the overlaid curve shows the prior predictions. Note that the elicitation is done predictively with respect to sets of covariates, where each covariate is given a unique value that affects the prediction. The figure is reproduced from James, Choy and Mengersen [42].

dictive distribution. Then, the hyperparameters of the chosen prior are fitted so that the expert's probabilities match with the prior predictive distribution. In practice, this

method becomes very complicated to implement when there are multiple parameters in the likelihood for which hyperparameters are being fitted [1].

A model-independent methodology for prior elicitation was introduced in the study of power priors [40, 41]. Power priors are subject-specific priors that are usually formed from a historical data y_0, although they arguably can also be constructed from expert opinions. In general form, power priors are given by

$$\pi(\boldsymbol{\theta}\,|y_0, \delta) \propto \pi(y_0|\,\boldsymbol{\theta})^{\delta}\pi_0(\boldsymbol{\theta}) \tag{2.6}$$

where is $0 \leq \delta \leq 1$ is a scalar parameter and $\pi_0(\boldsymbol{\theta})$ is the initial prior for the parameters $\boldsymbol{\theta}$ before the historical data y_0 are observed. Often, for $\pi_0(\boldsymbol{\theta})$ a reference prior is used. From above we see that δ controls the influence of the likelihood of historical data, thus it can be interpreted as a precision parameter which controls heaviness of the tails of the power prior. The parameter δ can be a fixed value, or it can also have its own prior probability, in which case a normalization of the power prior is necessary [63].

Power priors have a solid foundation to form subject-specific priors, but their use is motivated mainly by incorporating historical data. In situations where little or no historical data is available expert elicitation is still needed. Presumably, for power priors the likelihood function in equation above would be replaced with an expert's elicited probability distribution. Power priors include the uncertainty of experts judgements with the use of concentration parameter δ. However, power priors do not employ the predictive distributions, and as such would require further prior predictive checks for model validation.

The model-independence of predictive elicitation enables interesting possibilities for the Bayesian workflow. Winkler [87] proposes that assuming the expert provides reliable information in the elicitation process, this information can be used in two phases of the Bayesian workflow, either together or separately. First, the analyst can use predictive elicitation for n different observation models $\pi_{1,\ldots,n}(y_0|\,\boldsymbol{\theta})$ to obtain prior distributions that provide the best fit. Then, the analyst can compare which model would provide prior predictive values most closely resembling those of the expert, and choose that model for the subsequent Bayesian inference. Second, assuming the analyst has chosen the observation model $\pi(y_0|\,\boldsymbol{\theta})$, they can compare which family of prior distributions, given the obtained prior in each family, would provide the best fit of prior predictive values with respect to the expert's judgements.

Winkler [87] provides toy examples for his line of thought, but the use of predictive elicitation in Bayesian model building, particularly complementing the prior checking, has been a subject to only little previous research. Garthwaite and Dickey [29] presented a method for variable selection using predictive elicitation for normal linear model. However, predictive elicitation has been more often used to form poste-

rior models, which are then used for model comparison [39, 41]. The new method [37], which is introduced thoroughly in Chapter 4, is model-independent and considers the uncertainty of the judgements experts provide. Moreover, it has properties that make it suitable for prior checking and model comparison.

3. Probability Elicitation from Experts

The aim of expert elicitation is to quantify the uncertainty of expert's beliefs about unknown quantities. These uncertainties are expressed as subjective probabilities and as such, or by fitting them into a more suitable parametric probability distribution, they can be used in Bayesian modelling. Often in expert elicitation, the expert is assumed to be a human, in which case considering cognitive biases and human capabilities is important when conducting the elicitation process. In more general terms, the expert can be any entity that can provide probabilistic judgements on the topic of interest.

A number of methods for elicitation have been proposed that lead to a realistic representation of the expert's views. However, subjective probabilities are rarely perfect, and addressing that in a justified manner is important. Since probability is the right measure for uncertainty, a probabilistic approach should be taken also with the uncertainty about the accuracy of subjective probabilities. This is exactly what is done in the methodology introduced in Chapter 4.

An important topic in expert elicitation is combining probability distributions of multiple experts. Clemen and Winkler [15] categorize expert combination methods into mathematical and behavioural aggregations. The mathematical aggregation aims at producing a combined probability distribution using analytic processes, while in the behavioural aggregation, the experts interact together to produce a single consensus distribution. Naturally, the latter applies only to human experts whereas the former can be used in combining probability distributions regardless of their source. Nevertheless, these methods are not perfect and thus the uncertainty of aggregated probabilities should also be considered.

In this chapter, we will first briefly introduce a number of cognitive biases affecting the expert elicitation. Then, we discuss the general methods and considerations of practical expert elicitation. Finally, we go through different approaches proposed in previous research for combining multiple experts' judgements.

3.1 Cognitive aspects affecting elicitation

Since the pioneering research by Tversky and Kahneman [82] the effect of heuristics and biases in decision-making has become a widely studied area in psychology. Since the expert judgements are ultimately decisions, they are also prone to errors caused by psychological effects. Biases can be categorized into individual biases that affect an individual person, and group biases, that affect a group of people [4]. In this section, we review some of the more relevant biases in the expert elicitation.

Of the three biases and heuristics that Tversky and Kahneman introduce in [82], two are especially relevant in expert elicitation [60, 68]. These are availability and anchoring biases. According to the *availability bias*, people are prone to address higher probability to events they are readily able to recall. That is, events that are more salient are often attributed a higher subjective probability since they are easier to remember, regardless of their true probability. An common example is one where people tend to assign a relatively higher probability to a plane crash than to a car accident. This is likely due to the wider media coverage of the more extreme event. In elicitation, avoiding the availability bias is difficult, and it ultimately comes down to the expertise of the elicited subject.

Due to *anchoring bias* [82], people tend to initiate their answer from a readily available value and adjust to their judgement from that value. The initial value can be completely irrelevant to the task, but the judgement is still biased towards the anchoring value. In addition to being an individual bias, anchoring can also bias the decisions of a group, if the individual judgements are presented consecutively rather than concurrently. In this case, the latter judgements can be biased towards the first ones. Simple methods to decrease the anchoring bias include avoiding the use of default values and ensuring that experts' individual judgements are made independently from each other.

Another prevalent group bias along with the anchoring is the *herd behaviour* [4]. In herding, people tend to adjust their decisions towards the majority vote. A similar phenomena to herding is the *groupthink* [43], in which a group of people have a strong desire for conformity in making decisions. Hence, in combining multiple experts' judgements with the behavioral aggregation, these experts may neglect the unpopular views regardless of their logical coherence [53, 60, 68]. It is notable that although these are behavioural issues, mathematical aggregation may also be prone to errors caused by diminishing the less popular judgements. We will discuss this more in Section 3.3.

Overconfidence has also been noted as a major cognitive bias in expert elicitation [4, 53, 60, 68]. Particularly, overconfidence seems to affect the subjective probabilities close to 0 or 1. For example, people tend to assign the same probability when asked for

95% or 99% probability intervals. An interesting phenomenon related to overconfidence is the *hard-easy effect*, which states that a subject tends to be overconfident for hard but underconfident for easy questions [81]. The issues related to overconfidence should be considered when forming the elicitation questionnaire.

Other biases and heuristics that should be considered when forming the elicitation questionnaire include *representativeness heuristic* [82, 4, 53], *range-frequency heuristic* [71, 68] and the *framing effect* [83, 53]. Representativeness is described as the tendency to assign the probability that A belongs to B based on how similar A is to B. In a situation where A is very similar to B, the subject may attribute more probability to A belonging to B and C rather than A belonging to only C, even when this is less likely probabilistically. Following the range-frequency, people tend to assign probabilities evenly to the categories available. For example, if the subject is given four categories of events and the last is *all other events*, they are likely to address more probability on the first three events than when they are given seven possible events where the first three are the same as previously. The framing effect is related to how the problem is presented. As an example, when considering a purchase of tickets to a concert, an early-bird discount may sound more appealing than one with a late registration fee. In probability elicitation, questions about *lower 50% of the population* or *top 50% of the population* may yield different answers due to framing.

For further reading, Garthwaite, Kadane and O'Hagan [30] review expert elicitation methods in general and provide wide coverage on psychological considerations in elicitation. Baddeley, Curtis and Wood [4] introduce causes of biases in expert elicitation in the context of geosciences, Kynn [53] provides a general review of cognitive biases and heuristics affecting the expert elicitation and ten practical recommendations to minimize their effects, and Morgan [60] provides a critical discussion on the reliability of human experts due to the cognitive biases.

3.2 Single expert elicitation

Probability is found to be a proper measure for uncertainty. However, measuring probabilities is difficult, not only because of the biases discussed in the previous section, but also because it is genuinely hard to decide exact probabilities for events since in theory there should be a single probability value for each event. As O'Hagan and Oakley put it, "probability is perfect, but we can't elicit it perfectly." [70] To address this issue, they separate two challenges. First, we need to develop processes of elicitation that provide experts a way to express their uncertainties through probabilities. Second, we need to express the imprecision that inevitably remains regardless of the elicitation process. In this section, we will discuss these tasks from a single expert point of view.

3.2.1 Encoding uncertainty as probability

Encoding uncertainty as probability is the process where expert's knowledge is transformed into probability. Spetzler and Staël von Holstein [80] approach encoding from two angles based on what kind of questions the expert is asked, and what kind of responses are provided to those questions. The question types can be divided into direct and indirect types. In the *direct response mode*, the expert is asked to provide numeric answers to the questions. In the *indirect response mode*, the expert is asked to choose between two or more bets. In that case, the bets are adjusted until the expert is indifferent of the choice. Indirect questions can be further divided with regard to *external reference* and *internal* events. In the external reference process, one bet is set with respect to an uncertain quantity and another one is set with respect to a familiar reference event. When using the internal events, bets are defined to compare the likelihood of two or more uncertain events with each other.

Spetzler and Staël von Holstein [80] specify three basic types of encoding methods based on the responses for elicitation questions: P-methods acquire probability scales for fixed values of events; V-methods acquire value scales for fixed probabilities; and PV-methods acquire both scales jointly. All methods are expected to provide consistent judgments from the expert, that can be converted into a probability scale. However, asking the expert to provide moments of distribution or parameters of a known probability density function is not recommended because the full implications are rarely understood. A basic example of P-method would be asking the cumulative probability for events with different values (e.g. what is the probability that $A < 150$). Similarly, a basic example of V-method would be asking the values of events with different quantiles (e.g. what is the level of A given the lowest quartile). To illustrate, the `Elicitator` [42] uses quartiles as shown in Figure 3.1. Both of these examples represent the *direct response mode*. In [80] examples of encoding methods for all question and response types are given.

A special case of eliciting rare events is also discussed by Spetzler and Staël von Holstein [80]. Eliciting very small probabilities is generally difficult, since they are hard to discriminate for a human expert. In this case, the recommended approach is the indirect response mode with external reference events. For example, the expert could be asked to compare the probability of a rare poker hand with the uncertain event. Another method would be visualizing the probability of the event, for instance by coloring a portion of a large grid so that the colored area would correspond the probability. Further, other events could be colored on the same grid as well, so that the internal (other events) and external (the size of the grid) reference points would be visible for the expert.

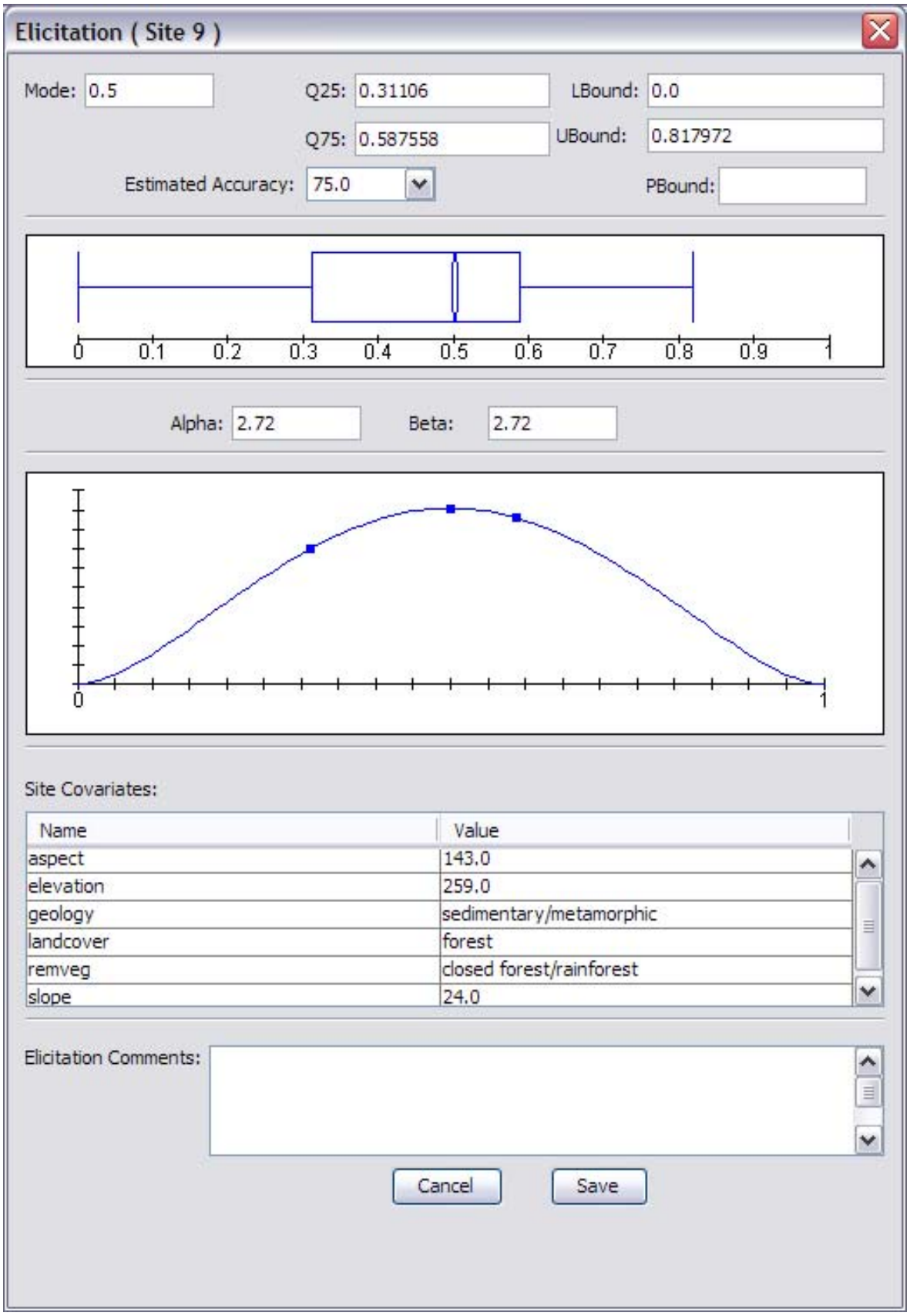


Figure 3.1: `Elicitator` [42] provides a V-method with direct response mode for probability elicitation given a set of covariates specified in the lower form. In the elicitation dialog, the expert can assign judgements by filling the values for quartiles. This also results in an interactive feedback in form of graphs. In addition, the expert can specify their confidence on the elicitation. The confidence or accuracy estimates are used as weights for judgements of different covariate sets in the predictive elicitation. The figure is reproduced from James, Choy and Mengersen [42].

In addition to the probabilities of events, elicitation of dependencies is often important. Clemen, Fischer and Winkler [14] assess six methods for pairwise dependency elicitation and find that, unlike recommended by Kadane and Wolfson [45], asking directly about correlation (which rarely is an observable quantity) seems to produce the best results. However, they advise to improve the accuracy of elicitation results by two methods. First, the outcomes should be visualized for the expert. This is in accordance with the desideratum of frequent feedback. For example, the `R` package `SHELF` [65] uses this method in forming bivariate distribution as shown in Figure 3.2. Second, the elicitation should be done with different methods, to examine the consistency of the expert's judgements. In [14] averaging of different elicitation outcomes is proposed, however, the expert could also simply reassess which of their judgements are the most reliable. These suggestions for improving the accuracy of elicitation are of course useful not only with the dependency elicitation but with the elicitation process in general.

3.2.2 Acknowledging imprecision in elicitation

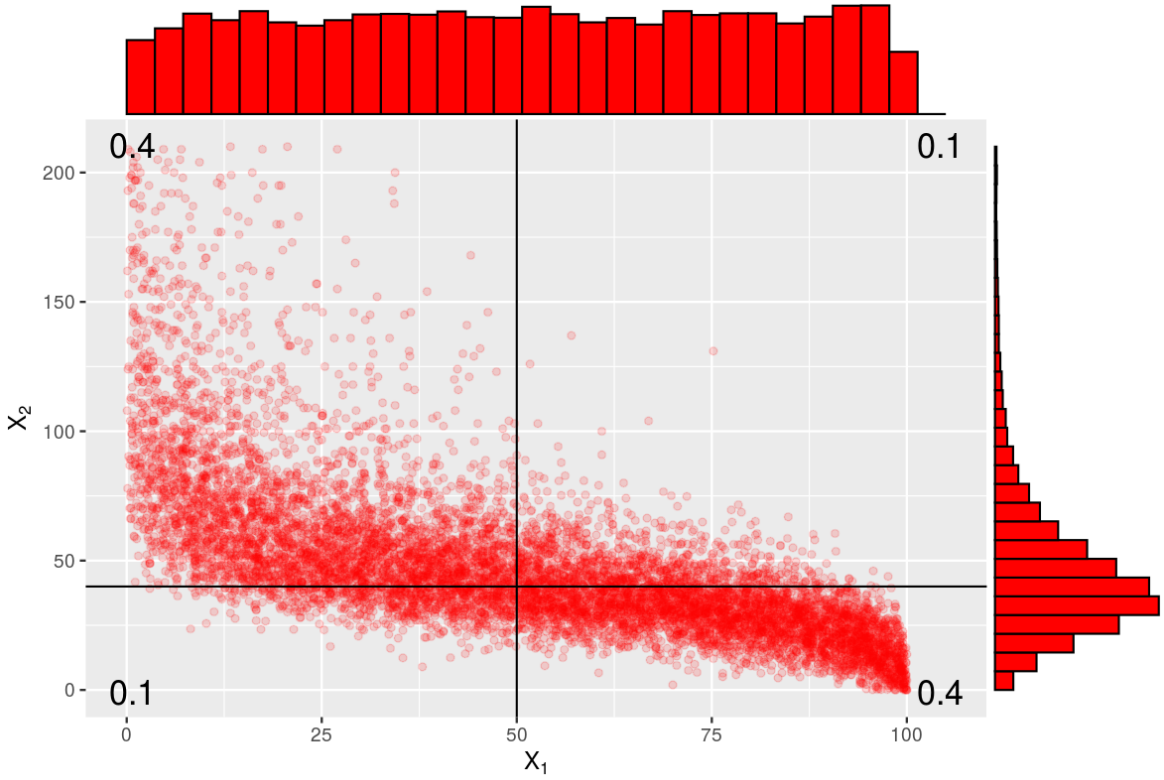
No matter how well the elicitation process is designed, there remains uncertainty about the precision of expert's knowledge. O'Hagan and Oakley [70] discuss a probabilistic approach to consider this uncertainty. They argue, that the analyst, who uses the expert's knowledge, has their own prior beliefs for the elicitation outcomes. A basic example of such prior beliefs would be that the probability density function being elicited is smooth and similar to a normal distribution. In a simple form, this can be done in practice by fitting a predefined parametric density function to the fractiles obtained in elicitation [67]. Generally, the idea is that the expert's judgements are regarded as stochastic realizations of their true knowledge.

Some practical tools have been developed for elicitation that also take the uncertainty of expert's precision into consideration. `Elicitator` [42] expands the idea of *conditional means* described in [7] to elicit predictively priors for GLMs. Unlike in conditional means, `Elicitator` allows the expert to provide judgements of outcomes on more covariate sets than the number of covariates, by formulating the elicitation process as a measurement error model. Thus, inconsistencies in the expert judgements are expected from a strict GLM point of view. However, the judgements are interpreted as outcomes of a conceptual model, which in turn is estimated by a regression model. The estimates are then used to find the prior distributions for coefficients of the GLM.

The Sheffield Elicitation Framework [35], and its `R` package implementation `SHELF` [65], is a tool for elicitation that consists of a software and an elicitation protocol, although here we focus on the former. The interface of `MATCH` [61], a web-based version

	L	0.25	0.5	0.75	U
Parameter 1	0	25	50	75	100
Parameter 2	0	30	40	60	200

(a) Elicited quartiles of the two parameters.



(b) Joint distribution of the best fitting distributions for the parameters 1 (x-axis) and 2 (y-axis). Plot shows 10 000 samples drawn from the joint distribution and the parameterwise histograms of those samples.

Figure 3.2: Elicitation of bivariate distribution is implemented in `SHELF` [65] by construction of the joint distribution using a Gaussian copula. The correlation parameter is determined by eliciting a concordance probability in the range $[0, 1]$, where probability > 0.5 implies positive correlation and probability < 0.5 implies negative correlation. **a)** In this example, we have two parameters for which we have elicit quartiles. According to `SHELF`, the best fitting distribution for the parameter 1 is Beta$(1, 1)$ and for the parameter 2 it is Log-Student's t with three degrees of freedom and $\mu = 3.72, \sigma = 0.456$. For this example, we choose a concordance $P(\{X_1 > m_1, X_2 > m_2\} \cup \{X_1 < m_1, X_2 < m_2\}) = 0.2$, where m_i is the median and X_i is a value of the parameter i. **b)** Finally, `SHELF` provides samples from the resulting joint distribution.

of `SHELF`, is shown in Figure 3.3. `SHELF` provides a number of parametric distributions that can be fitted to the judgements expert makes. It provides immediate feedback to the user in form of visualizations of elicited quantities as well as the fitted distributions. However, `SHELF` does not provide predictive elicitation for the Bayesian modelling.

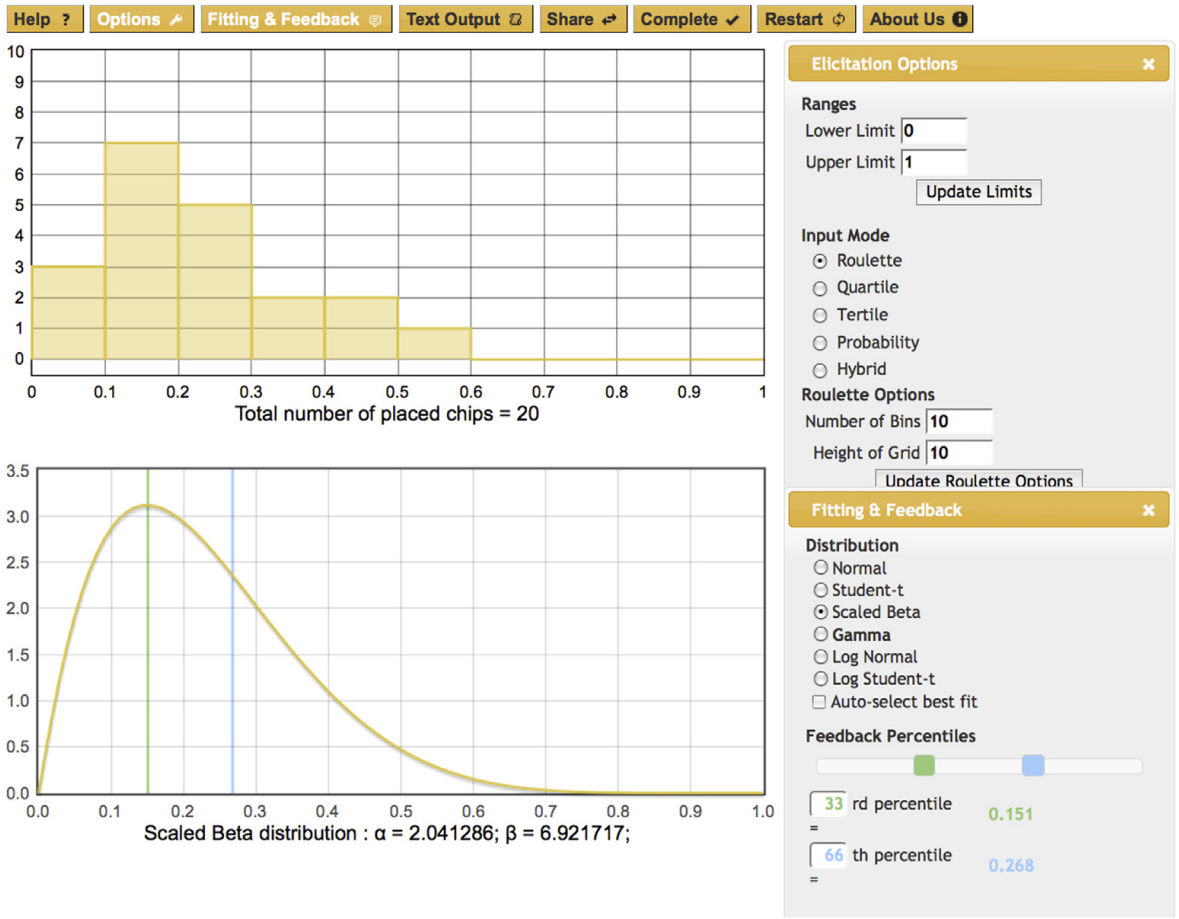


Figure 3.3: `MATCH` [61] is a web-based implementation of `SHELF` [65]. It provides P-, V- and PV-methods with direct response mode, and a V-method with indirect response for internal events called *trial roulette*, which is illustrated. In trial roulette, the expert places bets as chips into bins that represent alternative outcomes as shown in the upper graph. As with `SHELF`, the expert's judgements are fitted into a parametric distribution, in this case the beta distribution. This is shown in the lower graph as an immediate feedback for the user. The figure is reproduced from Morris, Oakley and Crowe [61].

3.3 Multiple experts

Combining probabilities of multiple experts usually begins with each expert providing their probabilities individually and independently, after which the analyst, or in this context the *facilitator*, either applies an analytic process to form a single probability distribution in the mathematical aggregation, or facilitates a structured interaction between the experts for them to arrive with a single consensus distribution. It should be noted that the aggregation method can also be a mix of mathematical and behavioural methods, such as the recently introduced IDEA protocol [38] and the repliCATS project [24] based on it, but the mixed methods are out of the scope of this thesis.

3.3.1 Mathematical aggregation

Clemen and Winkler [15, 16] identify two types of approaches for mathematical aggregation, axiomatic and Bayesian approaches. In axiomatic approaches, the strategy is to assume that the combined distribution should follow certain properties, and then the functional form of the combined distribution is derived. Popular examples of axiomatic approaches include *linear opinion pool* and *logarithmic opinion pool*. The formula of the linear opinion pool is given by

$$p(\theta) = \sum_{i=1}^{n} w_i p_i(\theta), \tag{3.1}$$

where n is the number of experts, $p_i(\theta)$ represents expert i's probability distribution for unknown θ, $p(\theta)$ represents the combined probability distribution, and the weights w_i are non-negative and sum to one. Thus, the linear opinion pool is just a weighted linear combination of experts' probabilities.

The formula of the logarithmic opinion pool is given by

$$p(\theta) = k \prod_{i=1}^{n} p_i(\theta)^{w_i}, \tag{3.2}$$

where k is a normalizing constant and the weights w_i satisfy some restrictions, such as they sum to one, to assure that $p(\theta)$ is a probability distribution. An important property of the logarithmic opinion pool is that it satisfies the principle of *external Bayesianity*. That is, if there is some new information about θ, it can be used to update individual experts' probabilities and then combine those again, or directly update the already combined probability distribution, which both will yield the same result.

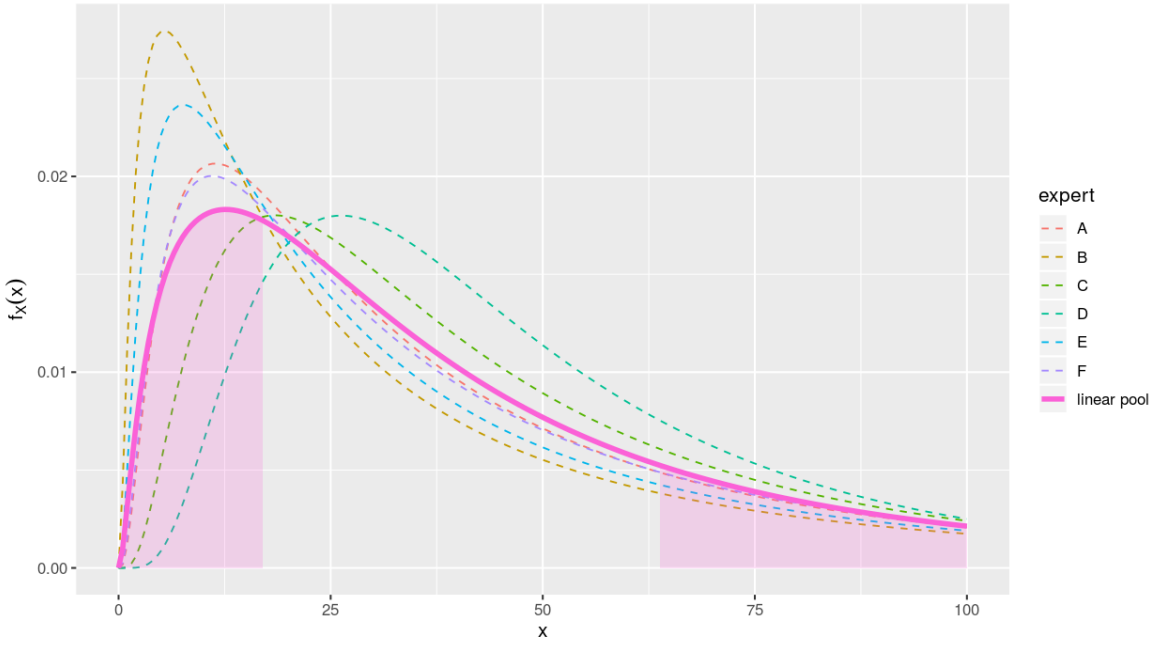
Bayesian approaches to aggregating expert information are based on using the Bayes' theorem to update prior distribution of the quantity of interest θ [15, 16]. Assuming that n experts provide their information $g_1, ..., g_n$ regarding θ, the facilitator will arrive with the updated probability distribution

$$p(\theta|g_1, ..., g_n) = \frac{L(g_1, ..., g_n|\theta)p(\theta)}{p(g_1, ..., g_n)} \tag{3.3}$$

where L is the likelihood function associated with the experts' information. L is central to the Bayesian aggregation, since it must account for the precision and bias of individual experts' information $g_1, ..., g_n$, and their mutual dependence. Here, the precision of g_i refers to the accuracy with which the expert i forecasts θ, and the bias of g_i is the extent to which the forecast consistently tends to fall with respect to θ. Clemen and Winkler [15, 16] reviewed a number of models designed to consider these interrelationships. Note that some of the models have been specifically designed to deal with issues yielding from cognitive biases, hence, allowing to account for biases not only in the individual elicitation process but also in the probability aggregation.

Expert	L	0.25	0.5	0.75	U	μ	σ
A	0	15	40	50	100	3.461	1.013
B	0	10	30	50	100	3.232	1.244
C	0	20	45	55	100	3.624	0.843
D	0	25	45	60	100	3.721	0.674
E	0	12	35	50	100	3.345	1.146
F	0	15	40	55	100	3.494	1.047

(a) Elicited quartiles of six experts and the parameters of the fitted log-normal distributions used in the linear pool. For all experts the lower bound is zero and the upper bound is one hundred.



(b) Densities of log-normal distributions fitted from experts' judgements. The shaded areas show the highest and lowest quartiles for feedback.

Figure 3.4: `SHELF` [65] provides option for linear pooling of multiple experts' judgements. **a)** The input of the judgements is numerical, either based on fractiles as used here or roulette bets. **b)** However, visual feedback is available, where either the fitted densities or histograms are plotted, including the distribution acquired by pooling.

3.3.2 Behavioural aggregation

In behavioral aggregation, experts arrive with a consensus probability distribution after a systematic interaction. Although prone to biases as discussed in Section 3.1, behavioural aggregation is still a common approach, particularly because it allows experts to debate their opinions [68]. A common behavioural aggregation method is

the Delphi method first introduced in [18], where in each round the experts make individual judgements, after which they are shared anonymously. Depending on the variation of the Delphi method, the experts then revise their probabilities and may have a critical discussion. After a few iterations, the experts arrive with an agreement or a mathematical aggregation is applied to combine their judgements [15, 68].

Behavioural aggregation approaches cover also the more informal methods, such as simply assembling the experts and assigning them the task of generating a single probability distribution. However, informal methods are prone to end with a disagreement or be affected by unhealthy group dynamics. Clemen and Winkler [15] list important issues when designing a behavioural aggregation. These include the type and the nature of interaction; the possibility of individual reassessment after interaction; and the role of the facilitator. In addition to improving the quality of the aggregation process, these are also important considerations when trying to diminish the effect of cognitive biases in behavioural aggregation.

3.3.3 Comparison of aggregation approaches

Mathematical and behavioural aggregation approaches both have their strengths and weaknesses. Behavioral aggregation is prone to psychological group effects and can be biased by those, whereas mathematical approaches avoid biases at the cost of losing the opportunity for the experts to debate their opinions [68]. A common problem that both approaches face is that the resulting combined probability distribution may not reflect anyone's particular probability distribution. In many mathematical approaches, this is self-evident, since they are often based on some aggregation procedure rather than choice of the best distribution. On the other hand, a behavioural aggregation may lead to a compromise. In both cases, the outcome can be based on contradictory information that leads to an unrealistic probability distribution.

Clemen and Winkler [15, 16] state that empirically the simple aggregation methods often outperform the more complex ones or at least reach a very similar level of success. They also emphasize that the cost of conducting a good behavioural aggregation can outweight the benefits of it. However, they do conclude that further work on the more complex Bayesian approach can lead to improved performance, since it allows formal adjustment for the quality of experts' judgements, including biases.

Regardless of which aggregation approach is used, the aggregated judgements include at least some uncertainty. This should be considered when, for example, performing prior elicitation for Bayesian modelling.

4. Probabilistic Predictive Elicitation

Our study [37], on which this thesis is based on, introduces a probabilistic method for predictive elicitation of priors, which provides an automatic transformation of expert's judgements about possible outcomes into subject-specific and generative priors of model parameters. Moreover, the methodology also allows the user to consider the uncertainty of correctness of the expert's opinions, which has not been previously considered in methodological literature to my knowledge. This general methodology is applicable to all model structures and is relatively simple to implement. In this thesis, we will refer to this method as the Probabilistic Predictive Elicitation (PPE).

In PPE the expert's judgements are required to imply mutually exclusive and collectively exhaustive set of probabilities for the possible outcomes of the generative process. The expert can express these probabilities in different ways as explained in Chapter 3, and the probabilities can be partitioned in whatever way the expert is comfortable with.

The fundamental idea of PPE is to assume that the expert's judgements are realizations of a Dirichlet distribution. This allows a probabilistic interpretation of the uncertainty in the expert's judgements, although here the prior elicitation is performed by maximizing the likelihood. The probabilistic interpretation of uncertainty in judgements is the main contribution of this methodology. For example, if the prior distributions would be elicited simply by matching the prior predictive probabilities with expert judgements, or as in [20] by matching the moments of prior predictive distribution with those elicited from the expert, the uncertainty of the expert's precision would not be accounted for.

In this chapter, we will first discuss important prerequisites regarding the mathematical machinery used. Then, we go through the basic model and the workflow of PPE. Finally, we analyze the distinctive properties of PPE. This chapter largely summarizes the description of PPE provided in our original work [37].

4.1 Prerequisites

Central to PPE are the prior predictive distribution (2.4) and the Dirichlet distribution, which is a multivariate generalization of the beta distribution. The prior predictive distribution was discussed in Chapter 2, and here it is important to note that in PPE the prior predictive distribution is assumed to express the possible outcomes given suitable prior parameters. In other words, in the standard use of PPE the model is expected to be properly chosen. In this section, we will first introduce main properties of the Dirichlet distribution, after which we formalize the prior predictive distribution for the purposes of PPE closely following the original paper [37].

4.1.1 Dirichlet distribution

Dirichlet distribution is a model of proportion variation. This subsection introduces the main properties of the Dirichlet distribution largely following the technical report by Minka [57].

Let $\mathbf{p}$ denote a random vector of length n whose elements sum to 1, and each element $\mathrm{p}_i > 0$. We assume that $\mathbf{p}$ follows the Dirichlet distribution $\mathcal{D}(\cdot)$ and denote the density at $\mathbf{p}$ as

$$\pi(\mathbf{p}) \sim \mathcal{D}(\mathbf{p}\,|\alpha, \mathbb{P}) = \frac{\Gamma(\alpha)}{\prod_{i=1}^{n} \Gamma(\alpha\ \mathbb{P})} \prod_{i=1}^{n} \mathrm{p}_i^{\alpha\,\mathbb{P}\,-1}, \tag{4.1}$$

where $\mathbb{P}$ is the mean vector of the distribution for $\mathbf{p}$ and α is a scalar that can be understood as a *precision* or *concentration* parameter.

The parameters α and $\mathbb{P}$ can be estimated given a training vector of proportions $\mathbf{p}$. The log-likelihood of the Dirichlet density function is written

$$\log\ \mathcal{D}(\mathbf{p}\,|\alpha, \mathbb{P}) = \log\ \Gamma(\alpha) - \sum_{i=1}^{n} \log\ \Gamma(\alpha\ \mathbb{P}) + \sum_{i=1}^{n} (\alpha\,\mathbb{P}\,-1) \log\ \mathrm{p}_i\,. \tag{4.2}$$

This objective function is convex in both α and $\mathbb{P}$ because the Dirichlet distribution is in the exponential family. Thus, the likelihood is unimodal and the maximum can be found by a simple search.

The maximum likelihood estimates (MLE) of mean $\mathbb{P}$ and precision α do not have a closed-form solution. However, Minka [57] provides a precise closed-form approximation for the MLE of α

$$\hat{\alpha} \approx \frac{n/2 - 1/2}{\mathrm{KL}(\mathbb{P}\,||\,\mathbf{p})} \tag{4.3}$$

where $\mathrm{KL}(\mathbb{P}\,||\,\mathbf{p})$ is the Kullback-Leibler divergence between the two distributions. The full derivation for the approximation is provided in Section 5.3. Furthermore, methods for finding MLE of $\mathbb{P}$ and α are discussed in Chapter 5 since they are central to PPE.

4.1.2 Prior predictive distribution in PPE

Recall the prior predictive distribution (2.4) of a Bayesian model $\mathcal{M}$. Let $\mathbf{y}$ be the a priori predicted realization of S-dimensional observable variables $\boldsymbol{Y} = [Y_1, ..., Y_S]$. Further, assume that the D-dimensional model parameters $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^D$ are distributed according to some prior distribution $\pi_{\boldsymbol{\theta}}$. Moreover, this prior $\pi_{\boldsymbol{\theta}}$ is assumed to belong to a family of parametric distributions $\mathcal{F}_{\boldsymbol{\lambda}}$ specified by a hyperparameter vector $\boldsymbol{\lambda}$. Since we assume that the Bayesian model $\mathcal{M}$ that we use in PPE is the correct model, it is ultimately described by the hyperparameters $\boldsymbol{\lambda}$. Therefore, we can rewrite the prior predictive distribution as

$$\pi_{\boldsymbol{Y}}(\mathbf{y} \,|\, \boldsymbol{\lambda}) = \int_{\Theta} \pi_{\boldsymbol{Y} \,|\, \boldsymbol{\theta}}(\mathbf{y} \,|\, \boldsymbol{\theta}, \boldsymbol{\lambda}) \pi_{\boldsymbol{\theta}}(\boldsymbol{\theta} \,|\, \boldsymbol{\lambda}) \mathrm{d}\boldsymbol{\theta} \,. \tag{4.4}$$

Denote the sample space Ω as a subset of $\mathbb{R}^S$ and let $\boldsymbol{A} = \{A_1, ..., A_n\} \subseteq \Omega$. In other words, $\boldsymbol{A}$ defines the *partitioning* of the sample space Ω. Throughout the elicitation procedure, the expert supplies their probabilistic judgements regarding the quantities $\mathbf{p}_{\boldsymbol{A}} = [\mathbf{p}_{A_1}, \ldots, \mathbf{p}_{A_n}]$, i.e. the subjective probabilities for all partitions. Moreover, the prior predictive distribution $\mathbb{P}(\boldsymbol{Y} \in A | \, \boldsymbol{\lambda}) = \mathbb{P}_{\mathrm{A} | \, \boldsymbol{\lambda}}$ of any partition $A \subseteq \Omega$ can be obtained by exchanging the order of integration via the Fubini-Tonelli theorem [23]:

$$\begin{aligned}
\mathbb{P}_{A | \, \boldsymbol{\lambda}} &:= \int_A \pi_{\boldsymbol{Y}}(\mathbf{y} \,|\, \boldsymbol{\lambda}) \, \mathrm{d}\mathbf{y} \\
&= \int_A \int_{\Theta} \pi_{\boldsymbol{Y} \,|\, \boldsymbol{\theta}}(\mathbf{y} \,|\, \boldsymbol{\theta}) \pi_{\boldsymbol{\theta}}(\boldsymbol{\theta} \,|\, \boldsymbol{\lambda}) \, \mathrm{d}\boldsymbol{\theta} \, \mathrm{d}\mathbf{y} \\
&\overset{\text{Fub.}}{=} \int_{\Theta} \int_A \pi_{\boldsymbol{Y} \,|\, \boldsymbol{\theta}}(\mathbf{y} \,|\, \boldsymbol{\theta}) \pi_{\boldsymbol{\theta}}(\boldsymbol{\theta} \,|\, \boldsymbol{\lambda}) \, \mathrm{d}\mathbf{y} \, \mathrm{d}\boldsymbol{\theta} \\
&= \int_{\Theta} \mathbb{P}_{\boldsymbol{Y} \,|\, \boldsymbol{\theta}}(\boldsymbol{Y} \in A | \, \boldsymbol{\theta}) \pi_{\boldsymbol{\theta}}(\boldsymbol{\theta} \,|\, \boldsymbol{\lambda}) \, \mathrm{d}\boldsymbol{\theta} \\
&= \mathbb{E}_{\boldsymbol{\theta}} \Big(\mathbb{P}_{\boldsymbol{Y} \,|\, \boldsymbol{\theta}}(\boldsymbol{Y} \in A | \, \boldsymbol{\theta}) \Big).
\end{aligned} \tag{4.5}$$

Note that the hyperparameter vector $\boldsymbol{\lambda}$, which defines a particular prior distribution $\pi_{\boldsymbol{\theta}}(\boldsymbol{\theta} \,|\, \boldsymbol{\lambda})$ from the set of all priors $\mathcal{F}_{\boldsymbol{\lambda}}$, is treated as constant. Indeed, in PPE the target is to find the value of $\boldsymbol{\lambda}$ that maximizes the Dirichlet likelihood given the expert's judgements $\mathbf{p}_{\boldsymbol{A}}$ for the partitions $\boldsymbol{A}$ and the corresponding prior predictive probabilities $\mathbb{P}_{\boldsymbol{A} \,|\, \boldsymbol{\lambda}} = [\mathbb{P}_{A_1 | \, \boldsymbol{\lambda}}, \ldots, \mathbb{P}_{A_n | \boldsymbol{\lambda}}]$.

Multidimensional data Recall that S denotes the dimensions of observable variables $\boldsymbol{Y}$. To partition the probabilities with an arbitrary number of dimensions S we can use a generic rectangular set $A = \bigtimes_{s=1}^{S} (a_s, b_s]$. Then, we can formulate the S-dimensional probabilities $\mathbb{P}_{A | \, \boldsymbol{\lambda}}$ using the cumulative distribution function of the prior predictive distribution (4.4).

Denote $F_{\boldsymbol{Y} \,|\, \boldsymbol{\lambda}}(\cdot)$ as the cumulative distribution function of the prior predictive distribution (4.4). Furthermore, consider an interval $I = (a, b]$, a function $g : \mathbb{R}^S \to \mathbb{R}$,

and a difference operator $\Delta_I^s = g(Y_1, \ldots, Y_{s-1}, b) - g(Y_1, \ldots, Y_{s-1}, a)$. Now, we can write the Equation (4.5) in a general form

$$\begin{aligned} \mathbb{P}_{A|\boldsymbol{\lambda}} &= \int_{a_1}^{b_1} \cdots \int_{a_S}^{b_S} \pi_{\boldsymbol{Y}\,|\,\boldsymbol{\lambda}}(Y_1, \ldots, Y_S) \mathrm{d}Y_1 \ldots \mathrm{d}Y_S \\ &= \Delta_{I_1}^1 \Delta_{I_2}^2 \cdots \Delta_{I_S}^S F_{\boldsymbol{Y}\,|\,\boldsymbol{\lambda}}(Y_1, \ldots, Y_S). \end{aligned} \tag{4.6}$$

This general formulation of *partitioned prior predictive distribution* is particularly useful for the computation of the Dirichlet MLE in Chapter 5.

4.2 The model

The assumption in PPE is that we have a Bayesian model that describes the data generating process, and we are interested in finding its prior probability distribution. Here, we assume that the parametric prior $\pi(\boldsymbol{\theta} \,|\, \boldsymbol{\lambda})$ is described by its unknown hyperparameters $\boldsymbol{\lambda}$. The expert is elicited probabilistic assignments $\mathbf{p}_{\boldsymbol{A}}$ regarding the data vector $\boldsymbol{Y}$. More precisely, these probabilities fall within a fixed set of mutually exclusive and exhaustive events $\boldsymbol{A}$. The assignments $\mathbf{p}_{\boldsymbol{A}}$ can be considered as the data for the inference which is conducted to retrieve the prior. However, it should be noted that $\mathbf{p}_{\boldsymbol{A}}$ is not to be confused with the actual measurement data from the generative process.

The core of PPE is the mathematical machinery that performs the predictive elicitation given the judgements of the expert and the Bayesian model. Therefore, this procedure is indifferent to how those judgements are collected, allowing a range of possible methods to be used as discussed in Chapter 3. Furthermore, regardless of the chosen method, PPE accounts for the uncertainty of the judgements with the use of probabilistic approach.

Following the description in the original work [37], the probabilistic predictive elicitation methodology of prior distribution for any Bayesian model is outlined as follows:

1. Define the probabilistic generative model for observable data $\boldsymbol{Y}$ conditioned on the parameters $\boldsymbol{\theta}$ and a parametric prior distribution $\pi_{\boldsymbol{\theta}}(\cdot)$ for those parameters. The prior distribution depends on hyperparameters $\boldsymbol{\lambda}$, which essentially define the prior and prior predictive distributions.

2. Partition the data space Ω of $\boldsymbol{Y}$ into exhaustive and mutually exclusive categories $\boldsymbol{A}$. For each of these categories, elicit from the expert the probability $\mathbf{p}_{A_j}$ that the outcome $\boldsymbol{Y}$ belongs to the category $A_j \in \boldsymbol{A}$.

3. Perform an iterative optimization of the Dirichlet MLE (4.1) with respect to $\boldsymbol{\lambda}$, where the support vector is $\mathbf{p}_{\boldsymbol{A}}$, and the Dirichlet mean is the partitioned prior predictive distribution $\mathbb{P}_{\boldsymbol{A}|\boldsymbol{\lambda}}$. The concentration parameter α can be fixed or estimated as well, as discussed in Section 4.3.1.

4. Evaluate how well the optimal prior predictive distribution found in Step 3 describes the elicited expert opinion.

One way to think about the process in PPE is that we try to find hyperparameters $\boldsymbol{\lambda}$ that maximize the matching S-dimensional histograms $\mathbb{P}_{\boldsymbol{A}|\boldsymbol{\lambda}}$ and $\mathbf{p}_{\boldsymbol{A}}$, with the constraint provided by the concentration parameter α in the Dirichlet likelihood function (4.1). When the histogram of prior predictive distribution is very similar to the histogram of the expert's probabilities, the $\hat{\alpha}$ which maximizes the likelihood is a large value. However, since α depends on the $\boldsymbol{\lambda}$ and vice versa, a great care should be taken in evaluating the results particularly if α is set to a fixed value.

Although none of the steps in PPE are trivial, guidance for implementation is provided in this thesis. Step 1 and Step 4 relate to each other in the sense that they require careful and possibly qualitative evaluation of the model selection. This was briefly touched in the end of Section 2.1 with references for further reading. The elicitation needed in the Step 2 plays an important part for the PPE to be successful, and considerations for it were discussed in Chapter 3. The optimization in the Step 3 is discussed in Chapter 5.

4.3 Properties

The novelty of PPE comes with how the uncertainty regarding the expert's judgements is dealt with. Moreover, PPE provides a flexible predictive elicitation framework which allows varying partitioning of probabilities to suite the expert's comfort zone. In addition, PPE is a model-independent method, which is novel in predictive elicitation literature. These properties are examined in this section by providing the theoretical background and practical examples from [37].

4.3.1 Uncertainty regarding the expert judgements

Probabilistic predictive elicitation differs from other predictive elicitation methods discussed in Section 2.3.2 in that it is truly model-independent while still taking into account the uncertainty in the expert's judgements. This uncertainty, that relates to the *imprecision* of the expert's judgments, is dealt with probabilistic approach by assuming that the probabilities assigned by the expert follow the Dirichlet distribution.

In particular, PPE uses the prior predictive probabilities $\mathbb{P}$ as the mean of the Dirichlet distribution (4.1) and the probabilities $\mathbf{p}$ elicited from the expert as the support of the Dirichlet distribution. Also, it is important to note that the prior predictive distribution must be partitioned to the same regime as the judgements given by the expert.

In PPE, we assume that the expert probabilities $\mathbf{p}$ are given and the probabilities $\mathbb{P}$ from the prior predictive distribution can be found. However, as shown in Chapter 5 the concentration parameter α is required for finding $\mathbb{P}$. Moreover, it becomes evident that α controls the uncertainty of expert's judgements given the model that produces $\mathbb{P}$. Thus, the analyst has two options, either find $\hat{\alpha}$ (together with the prior predictive distribution, hence prior) that maximizes the Dirichlet likelihood, or fix α to a predefined value that controls the variance in the Dirichlet distribution. The latter is a non-trivial approach because it means quantifying the uncertainty about the expert's knowledge a priori, but it simplifies the computation and may be of practical use.

The practical interpretation of the concentration parameter α stems from its approximation (4.3), particularly from the KL value in the denominator. With small KL values, the prior predictive probability could not be discriminated from the probabilities provided by the expert. The following example is provided in [37] to demonstrate the effect of different α values:

Example: Consider a generative model given by

$$y|\theta \sim \mathcal{N}(\theta, \sigma^2)$$

$$\theta \sim \tfrac{1}{2}\mathcal{N}(\mu_1, \sigma_1^2) + \tfrac{1}{2}\mathcal{N}(\mu_2, \sigma_2^2).$$

This yields the prior predictive distribution

$$y \sim \tfrac{1}{2}\mathcal{N}(\mu_1, \sigma^2 + \sigma_1^2) + \tfrac{1}{2}\mathcal{N}(\mu_2, \sigma^2 + \sigma_2^2)$$

with hyperparameters $\boldsymbol{\lambda} = [\mu_1,\ \mu_2,\ \sigma^2,\ \sigma_1^2,\ \sigma_2^2]^\top$.

For a set $A = (a, b] \subset \mathbb{R}$, the prior predictive probability is

$$\mathbb{P}_{A|\,\boldsymbol{\lambda}} = \sum_{k=1}^{2} \tfrac{1}{2}\Phi\Big((a-\mu_k)/\sqrt{\sigma^2+\sigma_k^2}\Big) - \tfrac{1}{2}\Phi\Big((b-\mu_k)\Big)/\sqrt{\sigma^2+\sigma_k^2}\Big).$$

Figure 4.1 illustrates the effect of the α parameter for a given partitioning $\mathbf{A}$ with $n = 10$. For each $\alpha \in \{1, 15, 50, 100, 300, 1000\}$, we generate $\mathbf{p}$ by sampling from (4.1), using fixed hyperparameter values of $\mu_1 = -\mu_2 = 2$ and $\sigma^2 = \sigma_1^2 = \sigma_2^2 = 1$.

In addition, the value of the concentration parameter α can be interpreted as the deviance between the prior predictive distribution and the expert's judgements. Thus,

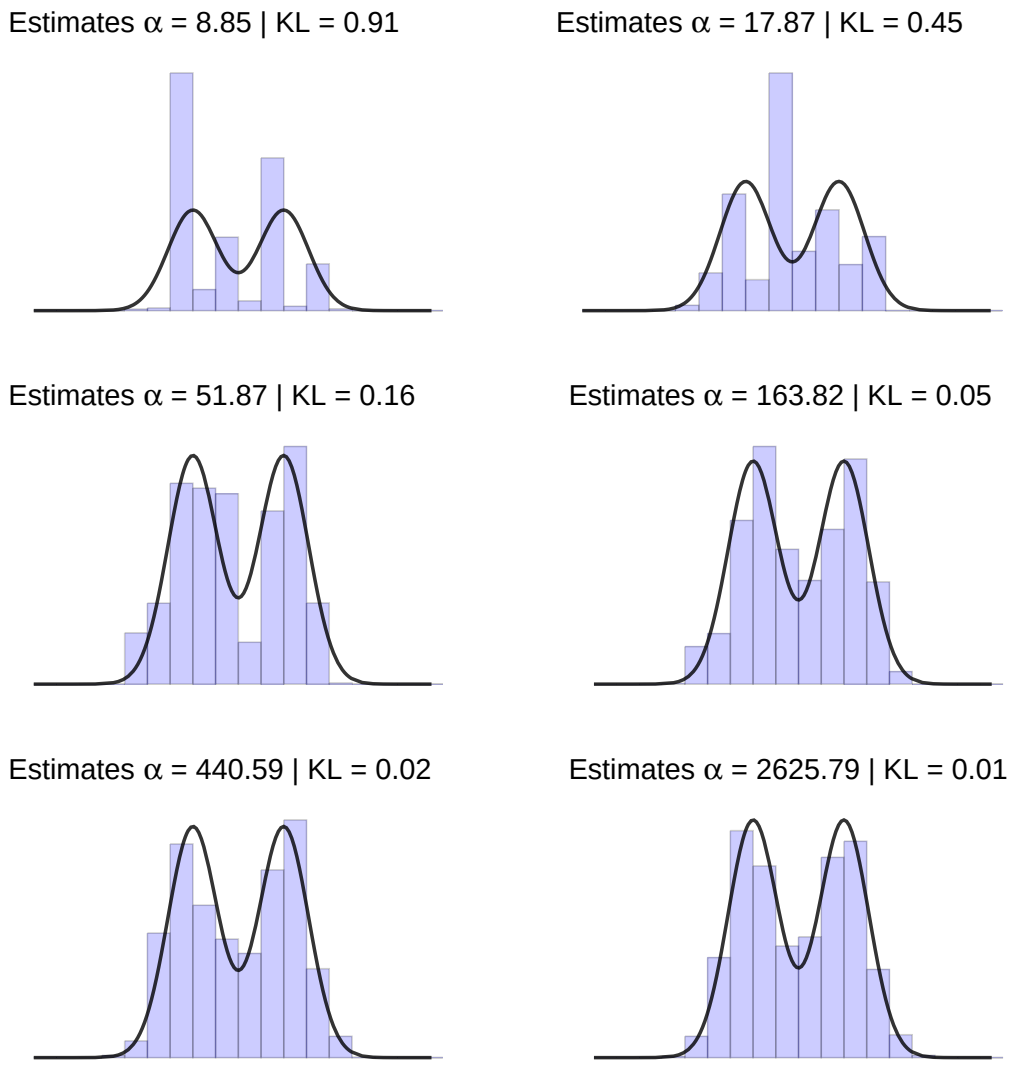


Figure 4.1: From [37]: Illustration of the role of the concentration parameter α. Large values correspond to scenarios where the prior predictive distribution (solid line) is able to represent expert's opinions (bars) accurately. That is, α provides an accuracy diagnostic for our method with higher values indicating higher accuracy.

if we are uncertain about the model choice, the expert's probabilities $\mathbf{p}$ can be used as the reference model and the MLE $\hat{\alpha}$ can be seen as the measure of model adequacy. Then, the model with larger $\hat{\alpha}$ implies that it is more adequate to express the data generating process according to the expert's beliefs. Yet, the stand-alone value of $\hat{\alpha}$ is ambiguous with the current knowledge, therefore it is recommended only for prior predictive model comparison, while predictive checks in form of visualization or other descriptive methods are encouraged for model validation.

4.3.2 Consistency with respect to partitioning

The PPE method allows the expert to provide their probabilities in partitioning of their choosing. Since PPE is inherently based on maximum likelihood estimation, the behaviour of MLE with respect to partitioning should be examined. The central underlying assumption is that the expert provides coherent probabilities regardless of the partitioning. That is, when the number of partitions n is increased, the expert provides more information about the probabilities, but they do not repeat the procedure of obtaining those probabilities multiple times. In this section we show that the MLE is consistent under this assumption and it results the true $\boldsymbol{\lambda}$ (with respect to the expert's judgements) when increasing n towards infinity.

In PPE the Equation (4.1) of the Dirichlet density represents the probabilistic model of $\mathbf{p}$ conditioned on the parameters $\boldsymbol{\tau} = (\boldsymbol{\lambda}, \alpha)$. Suppose that the expert's implied true prior distribution has hyperparameter values $\boldsymbol{\lambda}_0$ and denote $\boldsymbol{\tau}_0 = (\boldsymbol{\lambda}_0, \alpha_0)$, where α_0 maximizes the likelihood given $\boldsymbol{\lambda}_0$. Furthermore, assume a large partition size n and denote the log-likelihood as $T_{\boldsymbol{\tau}}(\mathbf{p}) = \log \mathcal{D}(\mathbf{p} \,|\alpha, \boldsymbol{\lambda})$ with the expectation $\mathbb{E}_{\mathcal{D}}[T_{\boldsymbol{\tau}}(\mathbf{p})]$ that is taken with respect to the Dirichlet distribution (4.1).

Now we can show that the expected Dirichlet log-likelihood is maximized at $\boldsymbol{\tau}_0$. Following the Jensen's inequality we know that

$$\mathbb{E}_{\mathcal{D}}\left[-\log \frac{\mathcal{D}(\mathbf{p} \,|\alpha, \boldsymbol{\lambda})}{\mathcal{D}(\mathbf{p} \,|\alpha_0, \boldsymbol{\lambda}_0)}\right] > -\log \mathbb{E}\left[\frac{\mathcal{D}(\mathbf{p} \,|\alpha, \boldsymbol{\lambda})}{\mathcal{D}(\mathbf{p} \,|\alpha_0, \boldsymbol{\lambda}_0)}\right]. \tag{4.7}$$

Furthermore, we note that the expectation on the right hand side of the equation equals one, thus the right hand side becomes zero and we can rewrite the inequality as

$$\mathbb{E}_{\mathcal{D}}\left[\log \mathcal{D}(\mathbf{p} \,|\alpha_0, \boldsymbol{\lambda}_0) - \log \mathcal{D}(\mathbf{p} \,|\alpha, \boldsymbol{\lambda})\right] > 0.$$

This yields

$$\mathbb{E}_{\mathcal{D}}[T_{\boldsymbol{\tau}_0}(\mathbf{p})] > \mathbb{E}_{\mathcal{D}}[T_{\boldsymbol{\tau}}(\mathbf{p})]$$

which holds for all $\boldsymbol{\tau}$, meaning that the expectation is maximized at $\boldsymbol{\tau}_0$.

The probabilistic model (4.1) must be identifiable to ensure the uniqueness of the MLE. That is, the equality of likelihoods must imply the equality of parameters: $\mathcal{D}(\mathbf{p} \,|\alpha_1, \boldsymbol{\lambda}_1) = \mathcal{D}(\mathbf{p} \,|\alpha_2, \boldsymbol{\lambda}_2) \Rightarrow \boldsymbol{\tau}_1 = \boldsymbol{\tau}_2$ for all $\mathbf{p}$. Otherwise, we may encounter multiple maxima and thus the elicited prior distribution in the set of possible priors $\mathcal{F}_{\boldsymbol{\lambda}}$ would not be unique. In practice, this may not be an issue when fitting the model since we are acquiring the prior with the assumption that further Bayesian analysis will be conducted with real observations. However, imposing identifiability does help avoiding problems in the optimization procedures discussed in Chapter 5.

The following example from [37] illustrates further the effect of more finely grained

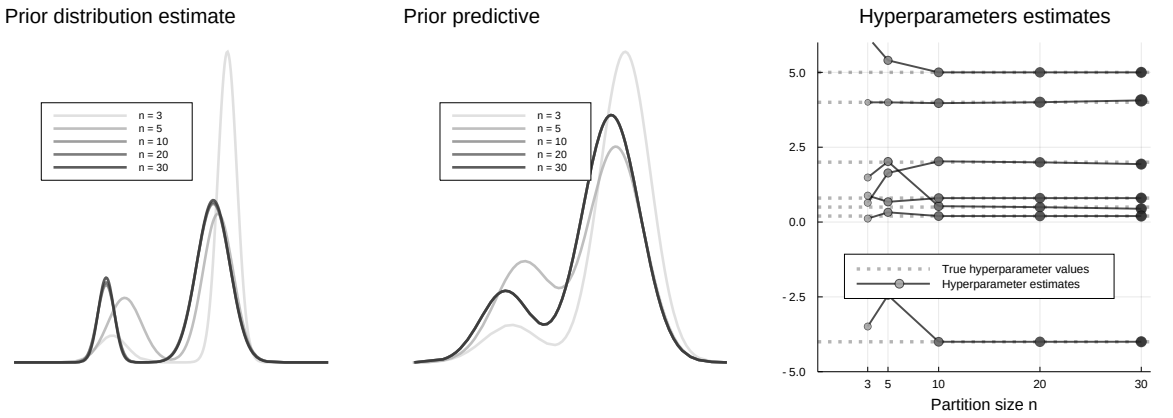


Figure 4.2: From [37]: Consistency of the MLE for $\boldsymbol{\lambda}$. **On the right:** All six hyperparameter values converge to the true values as the number of partitions n increases (each line corresponds to one hyperparameter), here converging already roughly for $n = 10$. **On the left and middle:** Both the estimated prior distribution (left) and the corresponding prior predictive distribution (middle) converge towards the respective true distributions, depicted as black lines.

partitioning for the expert probabilities. It also shows that even a small partition size n may provide suitable priors for the further inference task.

Example: Extending the earlier example, consider a more general generative model where the prior distribution is now

$$\theta \sim w_1\mathcal{N}(\mu_1, \sigma_1^2) + w_2\mathcal{N}(\mu_2, \sigma_2^2)$$

yielding the prior predictive distribution

$$y \sim w_1\mathcal{N}(\mu_1, \sigma^2 + \sigma_1^2) + w_2\mathcal{N}(\mu_2, \sigma^2 + \sigma_2^2),$$

where w_1 and w_2 are weights summing up to 1 and the hyperparameters are given by $\boldsymbol{\lambda} = [\mu_1, \mu_2, \sigma^2, \sigma_1^2, \sigma_2^2, w_1, w_2]$.

Suppose α is fixed and the true prior distribution has hyperparameters $\boldsymbol{\lambda}_0$. We run an experiment where probability vectors are generated from (4.1) with increasing partition sizes. Figure 4.2 shows that, as the partition size increases, the estimates $\hat{\boldsymbol{\lambda}}$ converge to $\boldsymbol{\lambda}_0$, which means the prior distribution is recovered from single-sample elicitation of probability data.

4.3.3 Covariate-dependent models

Models used in Bayesian inference are often covariate-dependent, i.e. there are some covariates $\mathbf{x}$ of which the outcome $\mathbf{y}$ are dependent. Following [37], here the procedure

of using PPE with covariate-dependent models is detailed for generalized linear models. Nevertheless, PPE is applicable for other types of covariate-dependent models as well.

GLMs consist of three elements: an exponential family of probability distributions; a linear predictor $\eta = \mathbf{x}\,\boldsymbol{\beta}$; and a link function g such that $\mathbb{E}(\mathbf{y}\,|\,\mathbf{x}) = \mu = g^{-1}(\eta)$. In a probabilistic approach, it is usually of interest to specify prior distributions for each parameter $\beta_c \in \boldsymbol{\beta}$ for C predicting covariates. In addition, priors may be wanted for an intercept of the linear predictor and potentially for a dispersion parameter depending on the chosen exponential distribution.

Specifying priors for covariate-dependent models is, however, often difficult. This is particularly true when the dependencies between parameters need to be specified by a joint prior distribution. Furthermore, even if the dependencies were not of interest, the implication of the parameters can be ambiguous as discussed in the context of structural elicitation in Section 2.3.1. Conversely, PPE can handle these issues elegantly.

In case of GLMs, the expert is required to provide judgements about plausible realization of $\boldsymbol{Y}$ given predefined covariate sets. For PPE, these covariate sets can be chosen such that the expert is comfortable to express probabilities for them*. Furthermore, the number of these covariate sets J is not fixed, and the partitioning $\boldsymbol{A}_j$ of the expert's probabilities $\mathbf{p}_j = \mathbf{p}_{\boldsymbol{Y}\,|\,\mathbf{x}_j}$ can vary by each covariate set $\mathbf{x}_j$. The latter provides more flexibility to the expert in expressing their knowledge of covariate-dependent data compared to alternative methods. For example, the conditional means method [7] requires the expert to provide a fixed number of probabilities for each covariate set to make the Jacobians used in the method invertible.

To formalize the above, we need to modify the equation of Dirichlet distribution (4.1) to account for J different probabilistic judgements. We start by expressing the procedure of eliciting expert judgements for covariate-dependent models by formal notation. First, we define the covariate sets $\mathbf{x}_j = [x_{j,1}, \ldots, x_{j,C}] \in \mathbf{X}$. Second, the expert provides probability judgements $\mathbf{p}_j = [p_{j,1}, \ldots, p_{j,n_j}]$ with $\sum_{i_j=1}^{n_j} p_{j,i_j} = 1$ for the outcomes of $\boldsymbol{Y}$ given the covariate sets $\mathbf{X}$. Here n_j is the size of partitioning $\mathbf{A}_j = [A_{j,1}, \ldots, A_{j,n_j}]$ of each judgement $\mathbf{p}_j$. Assuming that the expert judgements $\mathbf{p}_j \in \{\mathbf{p}_1, ..., \mathbf{p}_J\}$ for different covariate sets are pairwise conditionally independent, we can express the *modified Dirichlet likelihood* as a function of α and $\boldsymbol{\lambda}$

$$\mathcal{D}(\mathbf{p}_1, \ldots, \mathbf{p}_J\,|\alpha, \boldsymbol{\lambda}) = \frac{\Gamma(\alpha)^J}{\prod_{j=1}^{J}\prod_{i_j=1}^{n_j}\Gamma(\alpha\,\mathbb{P}_{A_{j,i_j}|\,\boldsymbol{\lambda},\mathbf{x}_j})}\prod_{j=1}^{J}\prod_{i_j=1}^{n_j} p_{j,i_j}^{\alpha\mathbb{P}_{A_{j,i_j}|\,\boldsymbol{\lambda},\mathbf{x}_j}-1}, \tag{4.8}$$

where $\mathbb{P}_{A_{j,i_j}|\,\boldsymbol{\lambda},\mathbf{x}_j}$ is the prior predictive probability for the set A_{j,i_j} given the covariate set $\mathbf{x}_j$. Note that the concentration parameter α of the Dirichlet likelihood remains a

*However, the choice of covariate sets can be in practice an influential decision for a successful elicitation. This is discussed in the context of a real experiment in Section 6.2

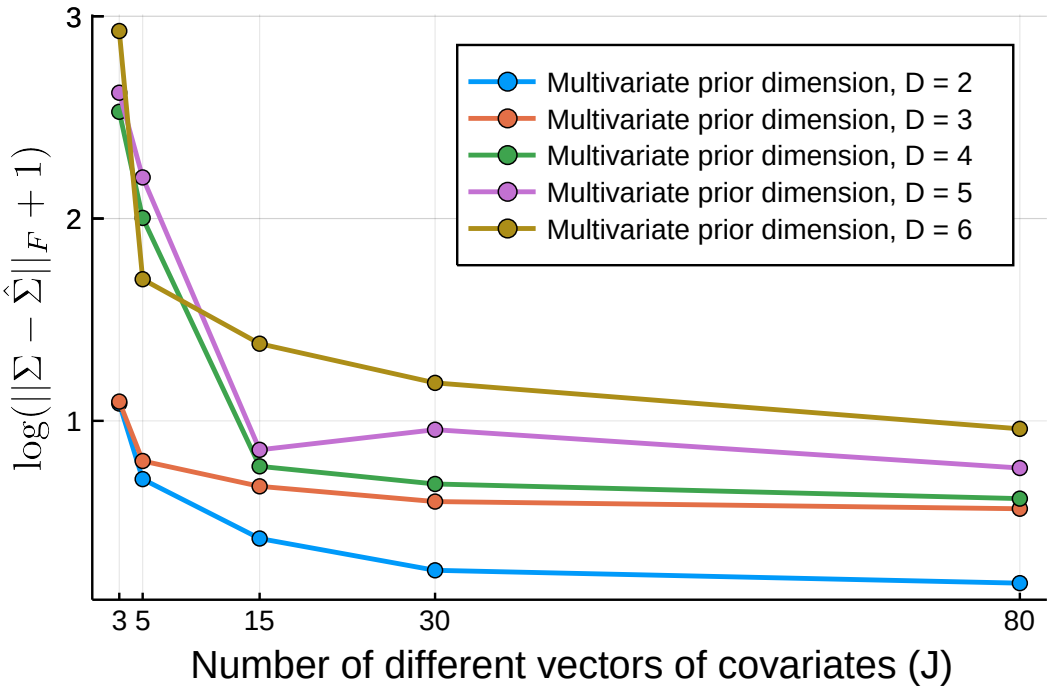


Figure 4.3: From [37]: Convergence of the covariance matrix estimates for multivariate prior elicitation for binary linear regression as a function of the number of covariates J for which the user provides probability estimates, measured using the logarithm of the Frobenius norm of the difference between the true covariance matrix and the estimate. The coloured lines refer to the dimensionality D of the prior distribution, showing that we can effectively elicit multivariate priors of reasonable dimensionality, with naturally increasing difficulty for larger D.

scalar value also in covariate-dependent models.

Although the number of covariate sets is not fixed in PPE, it is advised to produce expert judgements to more covariate sets as the size of hyperparameters $\boldsymbol{\lambda}$ that are elicited increases. The following example from [37] describes how the number of covariate sets affects the performance of elicitation:

> **Example:** Here we consider a generative model for binary data in the presence of a vector of covariates. The observable variable conditioned on the parameters is distributed according to a Bernoulli model and we take a multivariate Gaussian distribution as the prior distribution for the vector of parameters in the predictor function. This can be formalized as
>
> $$y|\,\boldsymbol{\theta} \sim \mathcal{B}\Big(\Phi(\mathbf{x}^\top \boldsymbol{\theta})\Big)$$
>
> $$\boldsymbol{\theta} \sim \mathcal{N}_D(\boldsymbol{\mu}, \Sigma)$$
>
> yielding the prior predictive distribution
>
> $$y \sim \mathcal{B}\Big(p(\mathbf{x}, \boldsymbol{\lambda})\Big)$$

with $p(\mathbf{x}, \boldsymbol{\lambda}) = \Phi(\mathbf{x}^\top \boldsymbol{\mu} / \sqrt{1 + \mathbf{x}^\top \Sigma \ \mathbf{x}})$.

The notation $\mathcal{N}_D(\cdot, \cdot)$ stands for a D-dimensional Gaussian distribution and $\mathcal{B}(\cdot)$ for the Bernoulli distribution. The hyperparameter vector $\boldsymbol{\lambda} = [\boldsymbol{\mu}, \Sigma]$, consists of the prior means $\boldsymbol{\mu} = [\mu_1, \cdots, \mu_D]$ and prior covariance matrix Σ. We fix the partitioning throughout the covariate set as $A_{j,1} = \{0\}$, $A_{j,2} = \{1\}$ since $y \in \Omega = \{0, 1\}$. Equation (4.5) simplifies to $\mathbb{P}_{A_1|\boldsymbol{\lambda}} = 1 - p(\mathbf{x}, \boldsymbol{\lambda})$ and $\mathbb{P}_{A_2|\boldsymbol{\lambda}} = p(\mathbf{x}, \boldsymbol{\lambda})$.

The parameterization of the covariance matrix follows the separation strategy suggested by [5] on an unconstrained space as presented by [52]. That is, the covariance matrix is rewritten as $\Sigma = \text{diag}(\sigma_1^2, \ldots, \sigma_D^2) \quad R \ \text{diag}(\sigma_1^2, \ldots, \sigma_D^2)$ where $(\sigma_1^2, \ldots, \sigma_D^2)$ are the variances and R is the correlation matrix.

In the simulation experiment, we vary the dimension $D \in \{2, 3, 4, 5, 6\}$ and the number of sets of covariates $J \in \{3, 5, 15, 30, 80\}$. For each D we randomly pick a true value for $\boldsymbol{\lambda}$, and for each covariate set, we draw random probabilities of success/failure from the Dirichlet probability model. Hence, the likelihood is given by (4.8). We repeat the procedure for each D and J where the hyperparameters $\boldsymbol{\lambda}$ are fixed with respect to J.

To show the convergence with respect to the estimates of Σ obtained from the expert judgements, we compare the logarithm of the Frobenius norm between the estimated covariance matrix and the true covariance matrix 4.3. For sufficiently large J, roughly from $J = 15$ onward, we are able to accurately elicit multivariate priors up to 5-6 dimensional priors – this is a significant improvement over earlier methods that have been limited to univariate or at most bivariate priors [58]. For increasing D from 2, 3, 4, 5 to 6, the respective number of hyperparameters in the vector $\boldsymbol{\lambda}$ becomes 5, 9, 14, 20 to 27, explaining the increased elicitation difficulty for large D.

5. Learning Methods

The computational difficulty in applying the PPE in practice comes with finding the hyperparameters $\boldsymbol{\lambda}$ that define the prior distribution, which in turn maximizes the Dirichlet likelihood through the prior predictive distribution. This requires an iterative optimization method, since there is no closed-form solution for the optimal mean vector of the Dirichlet distribution. Usually the formulation of PPE allows the use of gradient-based methods for optimization, which are recommended, but their practical implementation may be complicated.

In addition to finding the optimal hyperparameters $\boldsymbol{\lambda}$, the analyst may be interested in finding the optimal precision parameter $\hat{\alpha}$ for the Dirichlet distribution. Since the optimal $\boldsymbol{\lambda}$ depend on $\hat{\alpha}$ and vice versa, the optimization needs to be conducted simultaneously.

This chapter begins by providing a formulation for gradient-based learning with the assumption that the precision parameter α is fixed. Then, gradient-free approach is discussed. Finally, the methodology for finding $\hat{\alpha}$ is explained. Throughout this chapter we closely follow the description provided in our original work [37] and expand the discussion related to the computational methods.

5.1 Gradient-based learning

Gradient-based learning is used when optimizing convex (or concave) functions. First, the basic idea of such learning method is explained. Let $g(\boldsymbol{\lambda})$ be a convex function, and we want to find $\boldsymbol{\lambda}$ that minimizes the function iteratively. To do that we use *gradient descent*:
Choose an initial value $\boldsymbol{\lambda}_0$ for the parameters. Repeat until convergence:

1. Evaluate the gradient $\nabla g(\boldsymbol{\lambda}_i)$ at the current iteration i.
2. Update the parameters by moving to direction of the negative gradient: $\boldsymbol{\lambda}_{i+1} = \boldsymbol{\lambda}_i - \boldsymbol{\gamma}\, \nabla g(\boldsymbol{\lambda}_i)$.

The step-size $\boldsymbol{\gamma}$ controls the magnitude of the movement on each iteration. Various methods are proposed how to determine $\boldsymbol{\gamma}$, but they rely more on heuristics and em-

pirical results rather than on theoretical foundation. Thus, determining of $\boldsymbol{\gamma}$ is often problem-specific and not further discussed here.

In the context of PPE and Dirichlet MLE, the function we want to minimize is $-\mathcal{D}(\mathbf{p}\,|\,\mathbb{P}_{\boldsymbol{\lambda}}, \alpha)$ with respect to $\boldsymbol{\lambda}$. The assumption here is that the concentration parameter α is fixed. When we have a closed-form solution for the partitioned prior predictive distribution (4.6), we can use *automatic differentiation* to obtain $\nabla_{\boldsymbol{\lambda}}\,\mathbb{P}$ regardless of the chosen generative model. After that, we are left to find $\nabla_{\mathbb{P}}\,\mathcal{D}(\mathbf{p}\,|\,\mathbb{P}, \alpha)$. Given an unconstrained vector $\boldsymbol{z}$ of the same length n as $\mathbb{P}$, Minka proposes the following reparameterization to find the gradient of the log-likelihood (4.2):

$$\mathbb{P}_i = \frac{z_i}{\sum_j z_j} \tag{5.1}$$

$$\frac{\mathrm{d}z_i}{\mathrm{d}\,\mathbb{P}_i} = \sum_j z_j \tag{5.2}$$

$$\frac{\mathrm{d}\log\,\mathcal{D}(\mathbf{p}\,|\,\mathbb{P}_{\boldsymbol{\lambda}}, \alpha)}{\mathrm{d}z_i} = \frac{\alpha}{\sum_j z_j}\Big(\log\,\mathrm{p}_i - \psi(\alpha\,\mathbb{P}_i) - \sum_j^n \mathbb{P}_i(\log\,\mathrm{p}_j - \psi(\alpha\,\mathbb{P}_j))\Big), \tag{5.3}$$

where $\psi(\cdot)$ is the digamma function. Thus, we obtain the gradient $\nabla_{\mathbb{P}}\,\mathcal{D}(\mathbf{p}\,|\,\mathbb{P}, \alpha)$ and knowing $\nabla_{\boldsymbol{\lambda}}\,\mathbb{P}$ we easily compute the gradient $\nabla_{\boldsymbol{\lambda}}\,\mathcal{D}(\mathbf{p}\,|\,\mathbb{P}_{\boldsymbol{\lambda}}, \alpha)$ using the chain rule.

However, (4.6) is not always available in closed-form. In that case, we can use *implicit reparameterization gradient* [22] together with the automatic differentiation. That is discussed later in this section, but first we look at an important improvement to the plain gradient descent called *natural gradient* which is applicable here.

5.1.1 Natural gradients for closed-form cases

When the prior predictive distribution (4.6) is available in closed-form, any gradient-based optimization algorithm is applicable for computing the Dirichlet MLE. Natural gradient descent [3] is a particular technique, which provides fast convergence but requires computing the Fisher information matrix. In case of Dirichlet MLE, however, the Fisher information matrix for $\boldsymbol{\lambda}$ can be computed in closed-form. Since the Dirichlet distribution belongs the exponential family, the Fisher information for Dirichlet mean $\mathbb{P}$, or the partitioned prior predictive distribution, reads

$$H_{\mathbb{P}} = \alpha^2(\mathrm{diag}(\psi'(\alpha\mathbb{P})) - \psi'(\alpha)\mathbb{1}\mathbb{1}^\top), \tag{5.4}$$

where function $\psi'(\cdot)$ is the derivative of the digamma function. The inverse is given in closed-form as

$$H_{\mathbb{P}}^{-1} = \tfrac{1}{\alpha^2}\left(\mathrm{diag}(\psi'(\alpha\mathbb{P}))^{-1} + \frac{\mathrm{diag}(\psi'(\alpha\mathbb{P}))^{-1}\mathbb{1}\mathbb{1}^\top\,\mathrm{diag}(\psi'(\alpha\mathbb{P}))^{-1}}{(1/\psi'(\alpha) - \mathbb{1}^\top\,\mathrm{diag}(\psi'(\alpha\mathbb{P}))^{-1}\mathbb{1})}\right) \tag{5.5}$$

where $\mathbb{1}$ is $n \times 1$ vector with each component equals to 1.

Further, we need to obtain the Fisher information for the hyperparameters $\boldsymbol{\lambda}$. This can be done using the change of variables for a new parameterization, bypassing the need of recalculating integrals, as

$$H_{\boldsymbol{\lambda}} = (\nabla_{\boldsymbol{\lambda}} \mathbb{P})^\top H_{\mathbb{P}} (\nabla_{\boldsymbol{\lambda}} \mathbb{P}), \tag{5.6}$$

where each vector $\frac{d}{d\lambda_m}\mathbb{P} \in \nabla_{\boldsymbol{\lambda}} \mathbb{P}$ is a Jacobian matrix. Note that $H_{\mathbb{P}}$ is invertible and positive-definite. Thus, $H_{\boldsymbol{\lambda}}$ is also invertible and its Cholesky decomposition is stable to compute for inversion.

The natural gradient is obtained by

$$\tilde{\nabla}_{\boldsymbol{\lambda}} \mathcal{D}(\mathbf{p} \,|\, \mathbb{P}_{\boldsymbol{\lambda}}, \alpha) = H_{\boldsymbol{\lambda}}^{-1} \nabla_{\boldsymbol{\lambda}} Dir(\mathbf{p} \,|\, \mathbb{P}_{\boldsymbol{\lambda}}, \alpha), \tag{5.7}$$

which is then used in place of the regular gradient in the gradient descent algorithm above. Due to the closed-form expression, we can use natural gradients with almost no additional computational cost. This is important, because using natural gradient allows for fewer iterations, which is crucial when the cost of computation in each iteration is large. This is particularly true when (4.6) does not have a closed solution as discussed next.

Presence of covariates: Natural gradient is also applicable in PPE for covariate-dependent models. As discussed in Section 4.3.3, PPE accounts for different partitionings of probabilities in covariate-dependent models for each covariate sets. That is, for each covariate set $\mathbf{x}_j \in \mathbf{X}$ a different partitioning $\boldsymbol{A}_j = [A_{j,1}, ..., A_{j,n_j}]$ of the prior predictive probability $\mathbb{P}_{\boldsymbol{A}_j \,|\, \boldsymbol{\lambda}}$ is possible to use.

As shown in the Equation (4.8), the modified Dirichlet likelihood function for covariate-dependent models factorizes for distinct covariate sets. Therefore, we can simply sum the Fisher information matrices of $\mathbb{P}_{\boldsymbol{A}_j \,|\, \boldsymbol{\lambda}}$ for each covariate set to get the complete Fisher information [13]. Formally, this is written as

$$H_{\boldsymbol{\lambda}} = \sum_j \Big[(\nabla_{\boldsymbol{\lambda}} \mathbb{P}_j)^\top H_{\mathbb{P}_j} (\nabla_{\boldsymbol{\lambda}} \mathbb{P}_j) \Big], \tag{5.8}$$

where each $(\nabla_{\boldsymbol{\lambda}} \mathbb{P}_j)^\top H_{\mathbb{P}_j} (\nabla_{\boldsymbol{\lambda}} \mathbb{P}_j)$ is computed as discussed above.

5.1.2 Stochastic optimization

If the partitioned prior predictive distribution (4.6) cannot be expressed in closed-form but the Dirichlet likelihood (4.2) is differentiable with respect to $\boldsymbol{\lambda}$, we can use gradient-based optimization with reparameterization gradients and automatic differentiation.

The idea is that we can use Monte Carlo samples to estimate the elements of $\mathbb{P}_{\boldsymbol{A}\,|\,\boldsymbol{\lambda}}$ as well as the prior distribution $\pi(\boldsymbol{\theta}\,|\,\boldsymbol{\lambda})$ and we can perform automatic differentiation along the process. However, due to the stochasticity that the sampling provides, we cannot directly obtain the gradient $\nabla_{\boldsymbol{\lambda}}\,\mathbb{P}_{\boldsymbol{A}\,|\,\boldsymbol{\lambda}}$. Therefore, we use the reparameterization trick, where the goal is to find a *pivotal* or *standardization function* for the prior [13]. Then, we can obtain the gradient with a low computational cost [22, 59]. Similar approach has been used with prior predictive distribution in [20], where the moments of distribution were matched to find hyperparameters of a hierarchical model.

We start by noting that the partitioned prior predictive distribution, in other words the elements of vector $\mathbb{P}_{\boldsymbol{A}\,|\,\boldsymbol{\lambda}}$, are expected values with respect to the prior distribution $\pi(\boldsymbol{\theta})$. Hence, we can rewrite the partitioned prior predictive probability as an expected value

$$\mathbb{P}_{A|\,\boldsymbol{\lambda}} = \mathbb{E}_{\boldsymbol{\theta}}\left[\mathbb{P}(\boldsymbol{Y} \in A|\,\boldsymbol{\theta})\right] \tag{5.9}$$

which depends on hyperparameters $\boldsymbol{\lambda}$. Note however, that the expression $\mathbb{P}(\boldsymbol{Y} \in A|\,\boldsymbol{\theta})$ depends only on the parameters $\boldsymbol{\theta}$.

Next, we need to find a pivotal function $X = T(\boldsymbol{\theta})$ that removes the dependence of the sampled $\boldsymbol{\theta}$ on the hyperparameters $\boldsymbol{\lambda}$. In other words, it follows from the pivotal function that the distribution $\pi_X(\cdot)$ of X does not depend on $\boldsymbol{\lambda}$. Furthermore, the pivotal function X is required to be invertible and continuously differentiable with respect to its arguments and parameters. More precisely, the inverse in this case should be $\boldsymbol{\theta} = T_X^{-1}(\boldsymbol{\lambda})$.

We can then rewrite the expectation above with respect to the pivot X

$$\mathbb{P}_{A|\,\boldsymbol{\lambda}} = \mathbb{E}_X\left[\mathbb{P}(\boldsymbol{Y} \in A|T_X^{-1}(\boldsymbol{\lambda}))\right] \tag{5.10}$$

where the inverse function $T_X^{-1}(\cdot)$ depends on both, pivot X and hyperparameters $\boldsymbol{\lambda}$. The gradients can be computed interchanging the order of integration and derivation

$$\nabla_{\boldsymbol{\lambda}}\mathbb{P}_{A|\,\boldsymbol{\lambda}} = \mathbb{E}_X\left[\nabla_{\boldsymbol{\lambda}}\mathbb{P}(\boldsymbol{Y} \in A|T_X^{-1}(\boldsymbol{\lambda}))\right]. \tag{5.11}$$

Note that while X depends on $\boldsymbol{\theta}$, there is no need for resampling X because its distribution $\pi_X(\cdot)$ is not dependent on $\boldsymbol{\lambda}$ by definition of pivotal function. Thus, we can write the *explicit reparameterization gradient*, which is essentially an applied chain rule, as

$$\nabla_{\boldsymbol{\lambda}}\mathbb{P}_{A|\,\boldsymbol{\lambda}} = \mathbb{E}_X\left[\nabla_{\boldsymbol{\theta}}\,\mathbb{P}(\boldsymbol{Y} \in A|T_X^{-1}(\boldsymbol{\lambda}))\nabla_{\boldsymbol{\lambda}}T_X^{-1}(\boldsymbol{\lambda})\right]. \tag{5.12}$$

Furthermore, it is possible to avoid the inversion of the pivotal function by using the *implicit reparameterization gradient*. Following the Equation (6) from [22] we can write $\nabla_{\boldsymbol{\lambda}}T_X^{-1}(\boldsymbol{\lambda}) = -(\nabla_{\boldsymbol{\theta}}T(\boldsymbol{\theta}))^{-1}\,\nabla_{\boldsymbol{\lambda}}T(\boldsymbol{\theta})$. Now, the Equation (5.12) can be written as

$$\nabla_{\boldsymbol{\lambda}}\mathbb{P}_{A|\,\boldsymbol{\lambda}} = \mathbb{E}_X\left[\nabla_{\boldsymbol{\theta}}\,\mathbb{P}(\boldsymbol{Y} \in A|\,\boldsymbol{\theta}))\;\left(-(\nabla_{\boldsymbol{\theta}}T(\boldsymbol{\theta}))^{-1}\,\nabla_{\boldsymbol{\lambda}}T(\boldsymbol{\theta})\right)\right]. \tag{5.13}$$

For example, Figurnov et al. [22] suggest to use the cumulative distribution function (CDF) as the universal pivotal function, when it is assumed to be strictly monotonic and differentiable w.r.t. arguments and parameters (in this case sampled parameters $\boldsymbol{\theta}$ and hyperparameters $\boldsymbol{\lambda}$, respectively). With this assumption, we can write for univariate distributions $T(\boldsymbol{\theta}) = \Phi(\boldsymbol{\theta} \mid \boldsymbol{\lambda}) \sim \text{Uniform}(0, 1)$, and the gradient $\nabla_{\boldsymbol{\lambda}} \boldsymbol{\theta} = \nabla_{\boldsymbol{\lambda}} T_X^{-1}(\boldsymbol{\lambda})$ reads

$$\nabla_{\boldsymbol{\lambda}} T_X^{-1}(\boldsymbol{\lambda}) = -\frac{\nabla_{\boldsymbol{\lambda}} \Phi(\boldsymbol{\theta} \mid \boldsymbol{\lambda})}{\pi(\boldsymbol{\theta} \mid \boldsymbol{\lambda})}, \tag{5.14}$$

where $\Phi(\boldsymbol{\theta} \mid \boldsymbol{\lambda})$ is the CDF of prior distribution $\pi(\boldsymbol{\theta} \mid \boldsymbol{\lambda})$. The authors [22] also provide an extension of this universal pivotal function for the multivariate case. Using the implicit over the explicit reparameterization gradient enables easy and fast computation of the gradients for some standard distributions such as truncated, mixture, Gamma, Beta, Dirichlet, or von Mises [22].

5.1.3 Hierarchical models

Often a Bayesian model has a hierarchical structure which provides two key advantages [32]. First, it allows to utilize unobserved parameters $\boldsymbol{\theta}$ that affect the outcomes $\mathbf{y}$ by assigning them probabilistic distribution that can be described by other parameters. This is similar as discussed so far in this thesis, but the particular advantage is that we can add layers of parameters, both known and unknown, that together affect the outcome of the modelling. Second, hierarchical models are more appropriate for hierarchical data than simple nonhierarchical models. With such data, nonhierarchical models with few parameters lack the accuracy in fitting the data while nonhierarchical models with many parameters tend to overfit the training data. On the other hand, hierarchical models can have enough parameters to fit the data accurately while structuring the dependence of parameters to avoid overfitting. This section discusses how the natural gradient and reparameterization trick can be applied to hierarchical models in PPE.

We start by formally writing a hierarchical probabilistic model with L layers

$$\begin{aligned} \mathbf{y} \mid \boldsymbol{\theta}_1 &\sim \pi(\mathbf{y} \mid \boldsymbol{\theta}_1) \\ \boldsymbol{\theta}_1 \mid \boldsymbol{\theta}_2 &\sim \pi(\boldsymbol{\theta}_1 \mid \boldsymbol{\theta}_2) \\ &\vdots \\ \boldsymbol{\theta}_L \mid \boldsymbol{\lambda} &\sim \pi(\boldsymbol{\theta}_L \mid \boldsymbol{\lambda}). \end{aligned} \tag{5.15}$$

For this hierarchical model the prior predictive probability is

$$\begin{aligned}\mathbb{P}_{A|\boldsymbol{\lambda}} &= \int_{\Theta} \mathbb{P}(\boldsymbol{Y} \in A|\,\boldsymbol{\theta}_1) \prod_{\ell=1}^{L-1} \pi(\boldsymbol{\theta}_\ell \,|\, \boldsymbol{\theta}_{\ell+1}) \pi(\boldsymbol{\theta}_L \,|\, \boldsymbol{\lambda}) \mathrm{d}\boldsymbol{\theta} \\ &= \int_{\Theta_L} \pi(\boldsymbol{\theta}_L \,|\, \boldsymbol{\lambda}) \int_{\Theta_{L-1}} \pi(\boldsymbol{\theta}_{L-1} \,|\, \boldsymbol{\theta}_L) \cdots \int_{\Theta_1} \pi(\boldsymbol{\theta}_1 \,|\, \boldsymbol{\theta}_2) \mathbb{P}(\boldsymbol{Y} \in A|\,\boldsymbol{\theta}_1) \mathrm{d}\boldsymbol{\theta}_1 \ldots \mathrm{d}\boldsymbol{\theta}_L\end{aligned} \tag{5.16}$$

where $\Theta = \cup_{\ell=1}^{L} \Theta_j$ is the set of parameter spaces on all layers and $\boldsymbol{\theta}_\ell \in \Theta_\ell$ are the parameters on each layer. We can simplify the notation by writing the prior predictive distribution using the *tower property*. This is done by applying the expectations sequentially

$$\mathbb{P}_{A|\boldsymbol{\lambda}} = \mathbb{E}_{\boldsymbol{\theta}_L} \left(\mathbb{E}_{\boldsymbol{\theta}_{L-1}} \cdots (\mathbb{E}_{\boldsymbol{\theta}_1} (\mathbb{P}(\boldsymbol{Y} \in A|\,\boldsymbol{\theta}_1))) \right). \tag{5.17}$$

Hierarchical models often rely on sampling because of their complex structure. This naturally includes stochasticity in the model, and for gradient-based optimization the reparameterization gradient is useful as discussed in the previous section. To apply the reparameterization gradient, we need find a pivotal function $T_\ell(\boldsymbol{\theta}_\ell) = X_\ell \sim \pi_{X_\ell}(\cdot)$ for each layer ℓ whose inverse function is denoted as $\boldsymbol{\theta}_\ell = T_{X_\ell}^{-1}(\boldsymbol{\theta}_{\ell+1})$. Recall that since we assume a pivotal function for every layer ℓ, by definition the distribution of X, denoted as $\pi_{X_\ell}(X_\ell) = \pi_{\boldsymbol{\theta}_\ell \,|\, \boldsymbol{\theta}_{\ell+1}}(T_{X_\ell}^{-1}) |\det J(T_{X_\ell}^{-1})|$, does not dependent on any parameter $\boldsymbol{\theta}_{\ell+1}$ or the hyperparameter $\boldsymbol{\lambda}$. Thus, we define the composite of inverse functions for each layer as

$$\boldsymbol{\theta}_\ell = f_\ell(\boldsymbol{\lambda}) = (T_{X_\ell}^{-1} \circ T_{X_{\ell+1}}^{-1} \circ \cdots \circ T_{X_L}^{-1})(\boldsymbol{\lambda}). \tag{5.18}$$

With the reparameterization trick we can rewrite the expected value in Equation (5.17) as a function of $\boldsymbol{\lambda}$

$$\mathbb{P}_{A|\boldsymbol{\lambda}} = \mathbb{E}_{X_L} \left(\mathbb{E}_{X_{L-1}} \cdots (\mathbb{E}_{X_1} (\mathbb{P}(\boldsymbol{Y} \in A|f_1(\boldsymbol{\lambda})))) \right). \tag{5.19}$$

To get the gradient of the expectation (5.19) we need to start by conducting Monte Carlo sampling. Since $\boldsymbol{\lambda}$ is fixed, we first sample from $\pi(\boldsymbol{\theta}_L \,|\, \boldsymbol{\lambda})$ and layer by layer move towards $\pi(\mathbf{y} \,|\, \boldsymbol{\theta}_1)$. This way, we have acquired samples $f_\ell(\boldsymbol{\lambda})$ for each layer ℓ. To get the expectation, we simply calculate the sample mean of $\mathbb{P}_{A|f_1(\boldsymbol{\lambda})}$.

The gradient $\nabla_{\boldsymbol{\lambda}} \mathbb{P}_{A|\boldsymbol{\lambda}}$ is computed similarly to (5.12) using the reparameterization gradients on each layer where sampling is involved. We use the following expression to obtain partial derivatives

$$\frac{\mathrm{d}}{\mathrm{d}\lambda_m} \mathbb{P}_{A|\boldsymbol{\lambda}} = \mathbb{E}_{X_L} \left(\mathbb{E}_{X_{L-1}} \cdots \left(\mathbb{E}_{X_1} \left(\frac{\mathrm{d}f_1}{\mathrm{d}\lambda_m} \frac{\mathrm{d}}{\mathrm{d}f_1} \mathbb{P}_{A|f_1(\boldsymbol{\lambda})} \right) \right) \right), \tag{5.20}$$

where the first derivative on the right-hand side of the equation above then reads

$$\frac{\mathrm{d}f_1}{\mathrm{d}\lambda_m} = \prod_{\ell=1}^{L-1} \frac{\mathrm{d}T_{X_\ell}^{-1}}{\mathrm{d}T_{X_{\ell+1}}^{-1}} \frac{\mathrm{d}T_{X_L}^{-1}(\lambda_m)}{\mathrm{d}\lambda_m}. \tag{5.21}$$

That is, we compute the partial derivatives of the reparameterization gradient for each layer w.r.t. to the previous layer sampled, and apply the chain rule to combine them.

Similar to the previous section, when we cannot compute or otherwise want to avoid calculating the inverse function $T_{X_\ell^{-1}}$, we can proceed with the implicit reparameterization gradient. Here we derive the Equation (6) from [22] for the hierarchical model.

Since T_ℓ is a one-to-one function, we start by noting

$$X_\ell = T_\ell(T_{X_\ell}^{-1}(\boldsymbol{\theta}_{\ell+1})). \tag{5.22}$$

Similar to [22], we consider that the pivotal function $T_\ell(\boldsymbol{\theta}_\ell)$ depends by definition on the parameters $\boldsymbol{\theta}_{\ell+1}$ directly and also indirectly via the argument $\boldsymbol{\theta}_\ell$, while the sampled X_ℓ is independent of $\boldsymbol{\theta}_{\ell+1}$. Thus, we can utilize the *total derivative* with respect to $\boldsymbol{\theta}_{\ell+1}$ by using the implicit and explicit derivatives*

$$0 = \frac{\mathrm{d}T_\ell}{\mathrm{d}\,\boldsymbol{\theta}_{\ell+1}}\bigg|_{\text{explicit}} + \frac{\mathrm{d}T_\ell}{\mathrm{d}\,\boldsymbol{\theta}_{\ell+1}}\bigg|_{\text{implicit}} = \frac{\mathrm{d}T_\ell}{\mathrm{d}\,\boldsymbol{\theta}_{\ell+1}} + \frac{\mathrm{d}T_\ell}{\mathrm{d}\,\boldsymbol{\theta}_\ell}\frac{\mathrm{d}\,\boldsymbol{\theta}_\ell}{\mathrm{d}\,\boldsymbol{\theta}_{\ell+1}} \tag{5.23}$$

Identifying the notation $\boldsymbol{\theta}_\ell = T_{X_\ell}^{-1}$ for all ℓ and solving for $\dfrac{\mathrm{d}\,\boldsymbol{\theta}_\ell}{\mathrm{d}\,\boldsymbol{\theta}_{\ell+1}}$ yields the implicit reparameterization gradient for each layer ℓ

$$\frac{\mathrm{d}T_{X_\ell}^{-1}}{\mathrm{d}T_{X_{\ell+1}}^{-1}} = -\left(\frac{\mathrm{d}T_\ell}{\mathrm{d}T_{X_\ell}^{-1}}\right)^{-1} \frac{\mathrm{d}T_\ell}{\mathrm{d}T_{X_{\ell+1}}^{-1}}. \tag{5.24}$$

We can now plug Equation (5.24) into Equation (5.21) to use the implicit reparameterization gradient to estimate $\nabla_{\boldsymbol{\lambda}}\mathbb{P}_{A|\,\boldsymbol{\lambda}}$. Then, with the stochastic gradients we get the estimate for the Fisher information matrix $H_{\boldsymbol{\lambda}}$ as shown in equations (5.6) and (5.8). Finally, we can proceed with *stochastic natural gradient descent* to perform probabilistic predictive elicitation of the hyperparameters $\boldsymbol{\lambda}$ of the prior for general types of probabilistic models. The main advantage of this method is that it provides efficiency to the optimization by requiring few iterations.

*The terms *implicit* and *explicit* reparameterization gradients [22] come from the observation that the reparameterization gradient can be derived from the total gradient. Thus, the directly obtained reparameterization gradient is called explicit while the derived gradient is called explicit.

5.2 Gradient-free learning

The gradient-based learning for PPE discussed in the previous section is the recommended approach in finding $\boldsymbol{\lambda}$ due to its efficiency. However, it requires a number of computational procedures to work. This by itself may seem like a barrier of entry for PPE in an application. Moreover, in practice the gradient method can be difficult to implement when the model of interest is an arbitrary Bayesian network, and the implementation may be laborious with different prior distributions since they need to be differentiated. Therefore, gradient-free learning methods are applicable and useful to discuss in the context of PPE.

General-purpose global optimization tools, such as Bayesian optimization and Nelder-Mead, require only the ability to evaluate the objective (4.6) to determine optimal $\boldsymbol{\lambda}$. Furthermore, many practical optimization libraries (e.g. `optimR` [62]) provide extensive range of alternatives. In practice, these optimization methods work well with a relatively small number of hyperparameters $\boldsymbol{\lambda}$. However, when the objective (4.6) of an arbitrary model is evaluated by Monte Carlo sampling, the gradient-free methods face overhead from slow and stochastic evaluation. Therefore, a method such as Bayesian optimization, that only requires a few iterations and tolerates stochasticity is recommended.

5.3 Finding the concentration parameter α

Thus far, we have assumed that the Dirichlet precision parameter α has been fixed to some arbitrary value. Determining a suitable fixed value is non-trivial, and therefore the analyst may be inclined to find the parameter $\hat{\alpha}$ which maximizes the Dirichlet likelihood. Since the likelihood (4.2) depends on precision α and mean $\mathbb{P}$ (which in turn depends on $\boldsymbol{\lambda}$), the Dirichlet MLE is found by either alternating the optimization task of $\boldsymbol{\lambda}$ and α until convergence in case of gradient-based approach, or by combining the optimization task in case of gradient-free learning.

Minka [57] provides a gradient-based algorithm for finding α that is similar to Newton-Raphson method but has a faster convergence. However, the derivation of the method [56] is done for the basic case of Dirichlet distribution (4.1), so it does not account for different partitionings used in the covariate-dependent model (4.8). Addressing this issue provides an interesting topic for the further research, but in practice the currently available methods produce good results in the context of PPE.

Both Minka's and Newton-Raphson methods use the second derivative of target function. In case of Newton-Raphson for example, given a function $g(x)$ and its first

and second derivatives $g'(x)$ and $g''(x)$, we can find x that minimizes $g(x)$ by iteration

$$x_{i+1} = x_i - \frac{g'(x_i)}{g''(x_i)}. \tag{5.25}$$

For the Dirichlet log-likelihood (4.2), the first and second derivatives with respect to α are given by

$$\frac{\mathrm{d}}{\mathrm{d}\alpha} \log \mathcal{D} = \psi(\alpha) - \sum_{i=1}^{n} \mathbb{P}_i \, \psi(\alpha \, \mathbb{P}_i) + \sum_{i=1}^{n} \mathbb{P}_i \log \mathrm{p}_i \tag{5.26}$$

$$\frac{\mathrm{d}}{\mathrm{d}\alpha^2} \log \mathcal{D} = \psi'(\alpha) - \sum_{i=1}^{n} \mathbb{P}_i^2 \, \psi'(\alpha \, \mathbb{P}_i). \tag{5.27}$$

In case of covariate-dependent models in PPE, the derivatives are given according to the modified likelihood function (4.8) with J covariate sets

$$\frac{\mathrm{d}}{\mathrm{d}\alpha} \log \mathcal{D}_{\mathbf{p}_1,\ldots,\mathbf{p}_J} = J\psi(\alpha) - \sum_{j=1}^{J} \sum_{i=1}^{n_j} \mathbb{P}_{j,i} \, \psi(\alpha \, \mathbb{P}_{j,i}) + \sum_{j=1}^{J} \sum_{i=1}^{n_j} \mathbb{P}_{j,i} \log \mathrm{p}_{j,i} \tag{5.28}$$

$$\frac{\mathrm{d}}{\mathrm{d}\alpha^2} \log \mathcal{D}_{\mathbf{p}_1,\ldots,\mathbf{p}_J} = J\psi'(\alpha) - \sum_{j=1}^{J} \sum_{i=1}^{n_j} \mathbb{P}_{j,i}^2 \, \psi'(\alpha \, \mathbb{P}_{j,i}). \tag{5.29}$$

Minka [57] suggests using an approximation of α as a starting point for the iterative method. Here, we provide the derivation for a covariate-dependent Dirichlet likelihood. The Stirling's approximation of the $\Gamma(\cdot)$ function is given by

$$\Gamma(x) \approx \sqrt{\frac{2\pi}{x}} \left(\frac{x}{e}\right)^x. \tag{5.30}$$

We rewrite the likelihood function (4.8) with the above approximation and remove terms that do not depend on α for a simplified notation

$$\begin{aligned}
\mathcal{D}(\mathbf{p} \,|\alpha, \boldsymbol{\lambda}) &\approx \frac{\left(\sqrt{\frac{2\pi}{\alpha}} \left(\frac{\alpha}{e}\right)^{\alpha}\right)^{J}}{\prod\limits_{j,i_j} \sqrt{\frac{2\pi}{\alpha \mathbb{P}_{A_{j,i_j}|\boldsymbol{\lambda}}}} \left(\frac{\alpha \mathbb{P}_{A_{j,i_j}|\boldsymbol{\lambda}}}{e}\right)^{\alpha \mathbb{P}_{A_{j,i_j}|\boldsymbol{\lambda}}}} \exp\left(\sum_{i,j} \alpha(\mathbb{P}_{A_{j,i_j}|\boldsymbol{\lambda}} - 1) \log p_{j,i_j}\right) \\
&\approx \frac{\alpha^{\sum_j n_j/2 - J/2} \prod\limits_{i,j} \mathbb{P}^{1/2}_{A_{j,i_j}|\boldsymbol{\lambda}}}{\exp\left(\alpha \sum\limits_{i,j} \mathbb{P}_{A_{j,i_j}|\boldsymbol{\lambda}} \log \dfrac{\mathbb{P}_{A_{j,i_j}|\boldsymbol{\lambda}}}{p_{i,i_j}}\right)}.
\end{aligned} \tag{5.31}$$

The logarithm of the simplified equation above reads

$$\log \mathcal{D}(\mathbf{p}\,|\alpha, \boldsymbol{\lambda}) \approx \log\, \alpha \sum_j \frac{n_j - 1}{2} + \frac{1}{2} \sum_{i,j} \log\, \mathbb{P}_{A_{j,i_j}|\boldsymbol{\lambda}} - \alpha \sum_{i,j} \mathbb{P}_{A_{j,i_j}|\boldsymbol{\lambda}} \log\, \frac{\mathbb{P}_{A_{j,i_j}|\boldsymbol{\lambda}}}{p_{i,i_j}}. \tag{5.32}$$

Then, the $\hat{\alpha}$ that minimizes the approximated likelihood is given by

$$\hat{\alpha} \approx \frac{\sum_j \frac{n_j - 1}{2}}{\sum\limits_{i,j} \mathbb{P}_{A_{j,i_j}|\boldsymbol{\lambda}} \log\, \frac{\mathbb{P}_{A_{j,i_j}|\boldsymbol{\lambda}}}{p_{i,i_j}}} \approx \frac{\sum_j \frac{n_j - 1}{2}}{\sum\limits_j KL(\mathbb{P}_j || \,\mathbf{p}_j)}. \tag{5.33}$$

This approximation is particularly useful for PPE since it is accurate enough to be used alone and it reduces the time for computation when finding the $\boldsymbol{\lambda}$. However, when precise $\hat{\alpha}$ is required, for example in model comparison, the gradient-based method is preferred.

6. Experiments

To assess how probabilistic predictive elicitation behaves in practical applications we created an `R` language interface. To make the interface applicable to a variety of different models, we use a `brms` package [11] for Bayesian modelling. The package works as a wrapper for Bayesian generalized (non-)linear multivariate multilevel models using the probabilistic programming language `Stan` [12]. This approach allows flexible modelling with the caveat that the model outputs, including predictive distributions, are based on Markov chain Monte Carlo sampling. Therefore, in our applications the extraction of a predictive distribution is computationally heavy and the resulting distribution is stochastic.

For expert elicited probabilities, our interface uses the format used in `SHELF` package [65] for elicitation output. We implemented this utilizing the existing support for multiple expert elicitation, because we need to evaluate predictive probabilities for several covariates. Here, instead of reading the elicited values as probability sets by multiple experts, in our application we read them as probability sets for each covariate (set) by one expert. In other words, before the elicitation we define covariates or covariate sets for which the expert provides probabilistic assessments. Naturally this approach allows us, but does not oblige, to use the probability elicitation methods provided by the `SHELF` package.

The creation of the programming interface was an iterative process done concurrently with writing the conference paper [37]. Thus, the learning methods used in this chapter differ between the experiments. In the first two experiments, discussed in Section 6.1 and Section 6.2, we used gradient-free methods for optimization. However in the last experiment, in Section 6.3, we use a gradient-based method with *Adam* optimization algorithm [51]. The reason to use *Adam* was that it approximates the natural gradient well in a stochastic optimization problem. We did not use a stochastic natural gradient descent as its implementation proved to be difficult. The more thorough explanation about the interface is included in the Appendix A.

In this chapter, we use the `R` application in three experiments. In Section 6.1, we try to find similar predictive priors as in a previous research using PPE. This experiment was included in the first draft of the published paper [37]. In Section 6.2, which

is also in the published paper, we experiment and compare structural and predictive approaches to prior elicitation. In Section 6.3, we use PPE to conduct prior predictive model comparison and show the possibilities of model-independent approach to prior elicitation. The last experiment is an original study in this thesis. Through the experiments we find that PPE is suitable for all of the intended tasks.

6.1 Trauma center

In the first experiment, we want to see whether we can achieve similar results in predictive prior elicitation as in previous research. However, finding examples with sufficient expert probability data and elicited priors proved to be difficult. Bedrick, Christensen and Johnson [8] apply the idea of using conditional means [7] for prior specification in a Binomial regression. This method provides a fair point of comparison, since our programming interface was done to support probabilistic regression models. Although neither the elicited expert data nor the prior distributions were provided completely, we can arguably perform a sufficient comparison between the elicitation methods. We find that the PPE indeed is able to carry out a satisfactory predictive prior elicitation. This experiment was first discussed in the unpublished version of our original paper introducing PPE [37] and this section follows that description closely.

Following Bedrick, Christensen and Johnson [8], the task is to specify priors for a model that predicts whether a trauma center patient survives or not. The predictive variables are the patient's age (AGE) and the injury, which is described by three factors: the injury severity score (ISS) ranging from 0 (no injury) to 75 (severe injuries in three or more body areas); the revised trauma score (RTS) from 0 (no vital signs) to 7.84 (normal vital signs); and the predominant type of injury (TI) stating whether the injury is blunt ($TI = 0$) or penetrating ($TI = 1$). Further explanation of the factors is provided in the original study. In the predictive model, the scenarios are interpreted as covariate sets of a logistic regression, where the six-dimensional covariate vector is defined as $\mathbf{x} = [1, ISS, RTS, AGE, TI, AGE \times TI]$.

In the original study [8], a trauma surgeon was introduced with six different scenarios of injured patients. They were then asked to provide 1st, 50th and 99th percentiles for the probability of death in each of the scenarios. The elicited values are then fitted to six *Beta*-distributions describing the prior probability of death in each scenario. However, since the ultimate goal is to find priors for a binomial model, we are mainly interested in the expert's expectations. Therefore, we use the expected values of the beta distributions as a proxy of the expert probabilities for the patients' deaths in each scenario.

Similar to the original study [8], we use the logistic regression model and try to

	Mean		Std. dev.	
Variable	**PPE**	**BCJ**	**PPE**	**BCJ**
Intercept	-2.02	-1.79	2.08	1.10
ISS	0.08	0.07	0.03	0.02
RTS	-0.70	-0.60	0.16	0.14
AGE	0.04	0.05	0.00	0.01
TI	0.63	1.10	0.48	1.06
AGE × TI	-0.01	-0.02	0.00	0.03

Table 6.1: Comparison of the prior probability distributions elicited with probabilistic predictive elicitation (**PPE**) and the posterior means and standard deviations reported by Bedrick, Christensen and Johnson (**BCJ**) [8]. The variables represent covariates of the logistic model used in predicting the survival rates of trauma center patients.

fit the expert's probabilities of a patient's death with respect to each covariate set to obtain the parameters of the prior distribution. In our notation the model is written

$$Y \,|\, \boldsymbol{\theta} \sim \mathcal{B}\Big(p(\mathbf{x}, \boldsymbol{\theta})\Big)$$
$$\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\lambda})$$

with $p(\mathbf{x}, \boldsymbol{\theta}) = e^{\mathbf{x}^\top \boldsymbol{\theta}}/(1+e^{\mathbf{x}^\top \boldsymbol{\theta}})$ and hyperparameters $\boldsymbol{\lambda} = [\mu_1, \cdots, \mu_D, \mathrm{diag}[\sigma_1^2, \ldots, \sigma_D^2]$. Since the link function $p(\mathbf{x}, \boldsymbol{\theta})$ is given by the logistic function, the prior predictive distribution has no closed form. It reads

$$\pi_{Y \,|\, \boldsymbol{\lambda}, \mathbf{x}}(y) = \mathbb{E}_{\mathcal{N}(\boldsymbol{\theta} \,|\, \boldsymbol{\lambda})} \Big[p(\mathbf{x}, \boldsymbol{\theta})^y (1 - p(\mathbf{x}, \boldsymbol{\theta}))^{1-y} \Big]. \tag{6.1}$$

We assumed the parameters of the model to be independent to make our results more comparable to the original approach.

Table 6.1 compares our estimated results for mean and standard deviation parameters of each normally distributed parameter with the posterior means and standard deviations reported by [8]. Unfortunately, the original paper did not provide prior probability distributions. However, the study provided visual comparison of the prior and posterior probability distributions at different covariate sets, implying that the difference between prior and posterior distributions is not drastic. In this experiment we did not fix the concentration parameter α, and, we arrived with approximate α value of 50.9 for the optimal solution. This value is relatively high because Y is binary.

The results in the Table 6.1 show that we were able the replicate comparable priors as in [8]. We note similar means for all the covariates with a varying difference in the standard deviations when comparing our elicited prior distributions with the posteriors reported in the original study.

6.2 Human height growth

In our study [37], we conducted a small experiment with original expert data to evaluate the applicability of PPE in practice. The idea was to compare predictive and structural elicitation with a model that would require little prior expertise, since we did not want to recruit domain experts for a technical validation. Therefore, we chose to elicit prior probabilities for parameters of a widely used human height growth model with five parameters by Preece and Baines [77, Model 1]. The assumption was that everyone has some understanding of human height and thus can act as an expert. Our test subjects were five doctoral students in computer science with reasonable statistical knowledge. All participants reported that they were more comfortable providing probabilities for the predictive than the structural elicitation. Moreover, they were more confident about the priors elicited predictively to match their actual subjective priors than the priors elicited structurally. This section closely follows the description of the experiment given in the original paper [37] and adds to the discussion.

The human growth model [77] takes as inputs a time (age) covariate t and a five-dimensional parameter vector $\boldsymbol{\theta} = [\theta_1\ \theta_2\ \theta_3\ \theta_4\ \theta_5] = [h_1, h_{t_*}, t_*, s_0, s_1]^\top$, where h_1 is the average height of an adult human, h_{t_*} is the average height at a growth-spurt, t_* is the time (age) when that growth-spurt happens, and s_0 and s_1 are rate constants. With this notation the model reads

$$h(t;\boldsymbol{\theta}) = h_1 - \frac{2(h_1 - h_{t_*})}{e^{s_0\ (t-t_*)} + e^{s_1\ (t-t_*)}}. \tag{6.2}$$

The probabilistic model for observed data (human height) is specified as

$$\begin{aligned} Y_t|\,\boldsymbol{\theta}, b &\sim \mathcal{W}(h(t;\boldsymbol{\theta}), b) \\ b &\sim \mathcal{G}(a_0, b_0) \\ \theta_d &\overset{i.i.d}{\sim} \mathcal{LN}(a_d, b_d) \end{aligned} \tag{6.3}$$

where Y_t denotes the height of a human being at time t. $\mathcal{W}$, $\mathcal{G}$ and $\mathcal{LN}$ stand for Weibull, Gamma and log-Normal distributions, respectively. The scale parameter b controls the variance of the variable Y_t around $h(t;\boldsymbol{\theta})$: the larger the values of b the less variance around the $h(t;\boldsymbol{\theta})$ and vice versa. This description of the model and its parameters was also given to the participants.

The goal of the elicitation was to find for each expert the hyperparameter vector $\boldsymbol{\lambda} = \{a_m, b_m\}$, $m = 0, \ldots, 5$ that describes the expert's prior beliefs of the observed height Y_t at different ages t. For the purpose of clarity, we asked the participants to provide assessments of growth for a human male. The participants were provided with the following brief description of the human growth process and related general numerical values:

During the early stages of life the stature of female and male are about the same, but their stature start to clearly to differ during growth and in the later stages of life. In the early stage man and female are born roughly with the same stature, around 45cm - 55cm. By the time they are born reaching around 2.5 years old, both male and female present the highest growth rate (centimetres per year). It is the time they grow the fastest. During this period, man has higher growth rate compared to female. For both male and female there is a spurt growth in the pre-adulthood. For man, this phase shows fast growth rate varying in between 13-17 years old and female varying from 11-15. Also, male tend to keep growing with roughly constant rate until the age of 17-18, while female with until the age of 15-16. After this period of life they tend to establish their statures mostly around 162 - 190cm and 155 - 178cm respectively.

Given the background information and a description of the model, we performed structural and predictive elicitation with each participant. In the structural elicitation, the participants were asked to provide probabilistic assessments of the parameters $\boldsymbol{\theta}$ and b. This was done by asking the 10th, 25th, 50th, 75th and 90th percentiles of the possible values for each parameter. In the predictive elicitation, the participants were asked to provide distributions for human male heights at given ages $t = \{t_1, t_2, t_3, t_4\} = \{0, 2.5, 10, 17.5\}$. For this, we used the same percentiles as in the structural elicitation, thus obtaining the probabilities

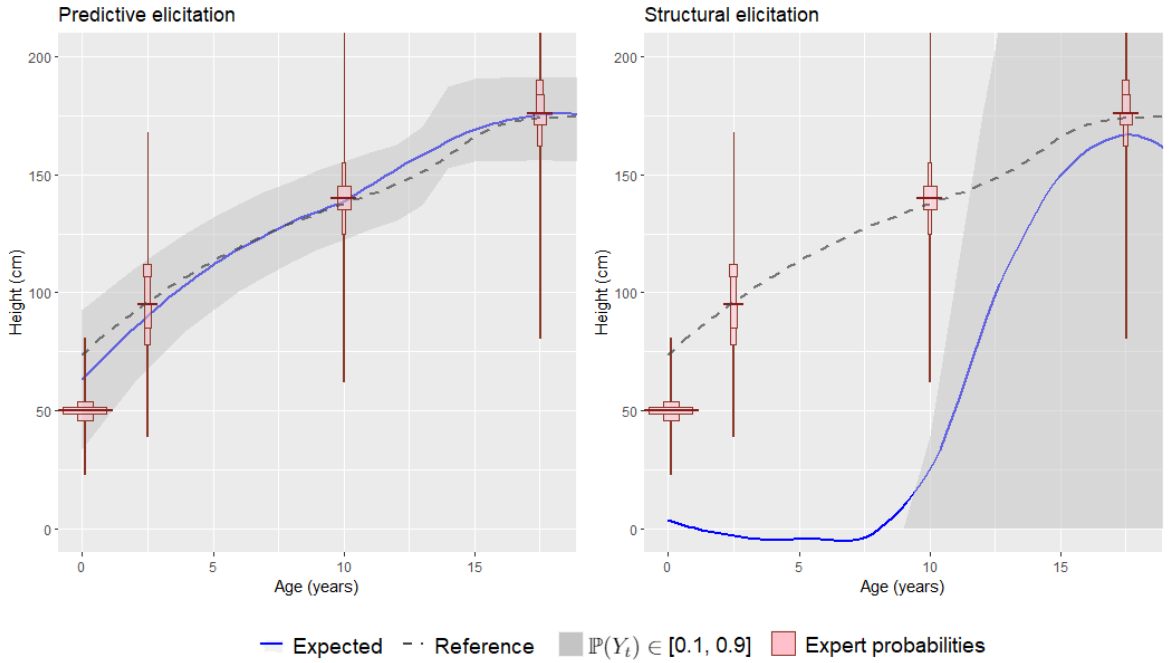
$$\begin{aligned}
\mathbb{P}(Y_t \leq y_1) &= 0.10 \\
\mathbb{P}(Y_t \leq y_2) &= 0.25 \\
\mathbb{P}(Y_t \leq y_3) &= 0.50 \\
\mathbb{P}(Y_t \leq y_4) &= 0.75 \\
\mathbb{P}(Y_t \leq y_5) &= 0.90
\end{aligned} \tag{6.4}$$

where naturally $y_1 < y_2 < \ldots < y_5$. The expert data used in PPE at each t_j was hence partitioned as

$$\begin{aligned}
\mathbb{P}(Y_{t_j} \in (0, y_1)) &= p_{j,i_j} = 0.10 \\
\mathbb{P}(Y_{t_j} \in (y_1, y_2)) &= p_{j,i_j} = 0.15 \\
\mathbb{P}(Y_{t_j} \in (y_2, y_3)) &= p_{j,i_j} = 0.25 \\
\mathbb{P}(Y_{t_j} \in (y_3, y_4)) &= p_{j,i_j} = 0.25 \\
\mathbb{P}(Y_{t_j} \in (y_4, y_5)) &= p_{j,i_j} = 0.15 \\
\mathbb{P}(Y_{t_j} \in (y_5, \infty)) &= p_{j,i_j} = 0.10.
\end{aligned} \tag{6.5}$$

		Predictive		**Structural**	
Parameter	Reference	$\mathbb{E}[\cdot]$	$\mathbb{V}(\cdot)$	$\mathbb{E}[\cdot]$	$\mathbb{V}(\cdot)$
h_1	174.6	174.5	0.8	176.2	105.3
h_{t_*}	162.9	162.8	4.2	129.1	33.6
s_0	0.1	0.1	< 0.1	1.2	1.1
s_1	1.2	3.3	0.2	1.2	1.1
t_*	14.6	13.4	0.01	12.5	0.6
b	–	15.8	12.9	2.0	4.6
α	–	6.9	–	1.2	–

(a) Elicited priors for one participant as shown in the original work [37]. For reference, the parameters presented by Preece and Baines [77] are included, except for the probabilistic model's parameter b. $\mathbb{E}[\cdot]$ is the sampled expected prior value for a parameter and $\mathbb{V}(\cdot)$ is the sampled prior variance.



(b) Prior elicitation results visualized for one participant. The plot on the right shows the results of predictive elicitation and on the left the structural elicitation. The blue line shows the sampled expectation of the prior predictive curve for different elicitation methods and the shaded area corresponds to the sampled prior predictive probability range $\mathbb{P}(Y_t) \in [0.1, 0.9]$. The vertical bars show expert probabilities proportionally to the area they cover, and the horizontal line on each bar plot shows the expert's expected height at the 50th percentile. For reference, the curve by Preece and Baines [77] is presented as a dashed line.

Figure 6.1: Elicitation results for one participant. The proposed approach (**Predictive**) achieves a better match for the expert expectations of height growth and the reference curve compared to the direct elicitation of parameters (**Structural**). Also, the lower α acquired in structural elicitation implies that the predictively elicited priors reflect better the expert expectations.

The results from structural elicitation had worse match with the participants' predictive expectations compared to the predictive elicitation. This was due to participants' inability to provide reasonable estimates for parameters in the structural elici-

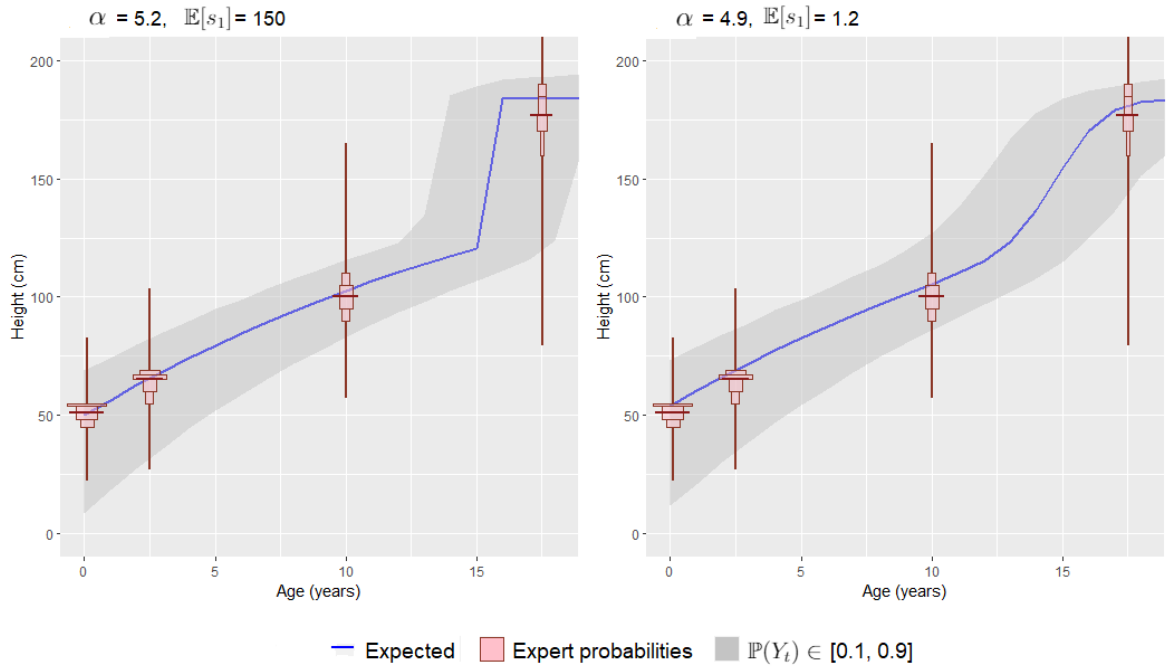


Figure 6.2: The importance of choosing covariate sets for elicitation. The plots show predictive elicitation results for one participant, when changing the prior probability for the rate parameter s_1. Since the change affects a covariate interval for which we have not elicited expert probabilities, the elicitation methodology is relatively ignorant about the parameter. Thus, the change has little effect on the concentration parameter α and the prior predictive probabilities at the elicited covariate locations.

tation despite being provided with an explanation of the model and its parameters. In a standardized interview after the elicitation, all participants told that they were more comfortable providing probabilities for human heights than model parameters. They were also more confident that the predictively elicited priors corresponded better their subjective prior views. This is in accordance with the Kadane and Wolfson's [45] elicitation desiderata. However, inconsistently with the desiderata, the participants were not provided with frequent feedback during the elicitation due to time constraints. Nevertheless, the lack of feedback applied to both elicitation methods making the results comparable with each other.

The Figure 6.1 shows the results of the two elicitation approaches with one participant. The plots show that the prior predictive growth curve elicited predictively has a much closer coverage of the expert's prior estimates of heights at different ages. Furthermore, the higher value of the concentration parameter α in predictively elicited priors shows that the prior predictive probabilities have a closer match to the expert probabilities. In this experiment we used the approximation of α, as it is sufficient to compare same models with different parameters. The elicited priors for all users are included in the Appendix B.

Although the priors obtained by PPE are well in line with the experts' expectations about the observable data, there is a caveat when conducting a predictive elicitation on covariate-based models. The covariate sets used for elicitation must be chosen carefully to cover features of the hypothetical observable data. Unfortunately, this was not taken properly into account in this experiment as the Figure 6.2 shows. For this participant, the significant change in prior probability of the rate constant s_1 has little effect on the concentration parameter α, but a visual assessment shows that the expected growth rate around the age 15 is too rapid. Moreover, in this case the Dirichlet likelihood that is being maximized is larger for the model with a larger expected rate s_1. This issue could have been easily resolved by adding the age covariate $t = 15$ in the elicitation. Despite this, the experiment shows that the predictive elicitation is a more appropriate approach to prior elicitation in probabilistic models than the structural elicitation. However, this does not decrease the importance of careful design of the elicitation process. To address this issue in future research, it would be beneficial to cover the automatic choice of covariates in the context of predictive prior elicitation.

6.3 Comparing height growth models

In the last experiment, we continue with the topic of human height growth. Here, we evaluate the suitability of PPE for prior model selection by comparing the behaviour of the concentration parameter α across different models for growth. Particularly, we assume that the models, which reflect the prior knowledge better, have a higher value of α. This is based on the observation we made in Section 4.1.1 that the denominator of the approximation of α is the KL-divergence between the model's prior predictive distribution and the predictive distribution provided by an expert. Thus, the higher the parameter α is, the less divergent the two distributions are. By comparing simple models with the more sophisticated growth models, we find that α is indicative of model adequacy. This experiment also emphasizes how predictive prior elicitation is model-independent by nature: because the expert provides predictive probabilities of observable quantities, the priors can be elicited for any model that is used to describe the generative process.

Following Ledford and Cole [54], the models we compare are the Preece and Baines model (PB) [77, Model 1] used in the previous section, the model by Jolicoeur, Pontier, Pernin and Sempé (JPPS) [44], and the Shohoji-Sasaki model modified by Cole (SSC) [17]. In addition, since all these models are designed to model the human height, we also compare the performance of PPE by modelling growth with a simple linear function (LINF) and a logistic function (LOGF). The linear model with slope and

intercept parameters is supposed to benchmark the least adequate human height growth model, because we expect the human growth to have some curvature due to growth spurts. Other models add complexity through the number of parameters: LOGF, PB, JPPS and SSC have three, five, seven and seven parameters, respectively. Naturally, all models also take age as a covariate.

For simplicity, we use the exact same probabilistic model as described in Equation (6.3) of the previous section for human height. Also, as mentioned before, the model PB, $h_{\mathrm{PB}}(t;\boldsymbol{\theta}_{\mathrm{PB}})$, is given in the Equation (6.2) in the previous section. The linear model LINF with parameters $\boldsymbol{\theta}_{\mathrm{LINF}} = [\beta_0, \beta_1]$ reads

$$h_{\mathrm{LINF}}(t;\boldsymbol{\theta}_{\mathrm{LINF}}) = \beta_0 + \beta_1 t, \tag{6.6}$$

where β_0 is the intercept and β_1 is the slope. The logistic model LOGF has three parameters $\boldsymbol{\theta}_{\mathrm{LOGF}} = [t_0, h_1, k]$ and it is given by

$$h_{\mathrm{LOGF}}(t;\boldsymbol{\theta}_{\mathrm{LOGF}}) = \frac{h_1}{1 + e^{-k(t-t_0)}}, \tag{6.7}$$

where t_0 is the newborn age, h_1 is the height at the maturity, and k is the logistic growth rate.

Unlike the other models, the model JPPS covers the height growth from conception. Therefore, to use the same age covariate as in the rest of the models, we use an adjustment term $t' = t+0.75$ assuming a constant gestation of nine months. Furthermore, JPPS has a seven-dimensional parameter vector $\boldsymbol{\theta}_{\mathrm{JPPS}} = [h_1, C_1, C_2, C_3, D_1, D_2, D_3]$ and the model is written as

$$h_{\mathrm{JPPS}}(t;\boldsymbol{\theta}_{\mathrm{JPPS}}) = h_1 \left(1 - \frac{1}{1 + (t'/D_1)^{C_1} + (t'/D_2)^{C_2} + (t'/D_3)^{C_3}}\right), \tag{6.8}$$

where h_1 is the height at the maturity, D_1, D_2 and D_3 are positive age scale factors, and C_1, C_2 and C_3 are positive dimensionless exponents.

Similar to JPPS, the model SSC has a seven-dimensional parameter vector $\boldsymbol{\theta}_{\mathrm{SSC}} = [h_1, k, \beta_0, \beta_1, c, r, t_*]$. SSC is described as a combination of an exponential infancy, linear childhood and logistic puberty components and reads

$$h_{\mathrm{SSC}}(t;\boldsymbol{\theta}_{\mathrm{SSC}}) = 0.1\Big(h_1 W(t) + f(t)[1 - W(t)]\Big), \tag{6.9}$$

where h_1 is again the height at the maturity, $f(t)$ is a function of height in infancy, and $W(t)$ is a weighting function. Thus, height at age t is a weighted average of adult height h_1 and the predicted height of early childhood $f(t)$. For childhood components, SSC uses The Jenss-Bayley function $f(t) = \beta_0 + \beta_1 t - e^{c-rt}$, which combines linear and exponential growth components. The parameter r controls the length of the infancy growth spurt while the constant c controls the height at birth. The intercept β_0 controls

$\mathbb{P}(Y_t \leq y_t)$	$y_{t=0}$	$y_{t=5}$	$y_{t=10}$	$y_{t=12}$	$y_{t=15}$	$y_{t=20}$
0.00	40	98	125	133	151	162
0.025	46	103	130	138	156	167
0.16	48	107	135	145	164	173
0.50	50	112	141	152	172	180
0.84	52	116	148	159	180	186
0.975	54	120	153	167	188	192
1.00	60	125	158	172	193	197

Table 6.2: Expert data at different ages used for predictive prior elicitation on different growth models. The leftmost column $\mathbb{P}(Y_t \leq y_t)$ shows the cumulative probability of height Y_t being lower than the height y_t indicated by the expert at the age t. The probabilities are based on a reported growth curve of Finnish boys [84].

childhood height level while slope β_1 controls the childhood growth. The weighting function $W(t) = e^{-e^{k(t_*-t)}}$ is the Gompertz function, a type of sigmoid function, which is zero at $t = 0$ and switches from 0 to 1, being e^{-1} at the age t_*, while the parameter k controls the suddenness of the switch. Additionally, the result of h_{SSC} is multiplied by 0.1 simply to convert millimetres to centimetres.

For this experiment, we simulate expert probabilities by gathering data from the real growth curve of Finnish boys reported by a research group in the University of Eastern Finland and Kuopio University Hospital [84]. The curves provide average heights at different ages, and the heights at one and two standard deviations away from the average. For simplicity, we assume the data to be normally distributed and arbitrarily round the probabilities at standard deviations, thus, obtaining probabilities

$$\begin{aligned}
\mathbb{P}(Y_t \leq y_1) &= 0.025 \\
\mathbb{P}(Y_t \leq y_2) &= 0.160 \\
\mathbb{P}(Y_t \leq y_3) &= 0.500 \\
\mathbb{P}(Y_t \leq y_4) &= 0.840 \\
\mathbb{P}(Y_t \leq y_5) &= 0.975
\end{aligned} \tag{6.10}$$

where similar to previous section $y_1 < y_2 < \ldots < y_5$. To simplify the computation, we limit the maximum (minimum) height at each age to height at the second standard deviation plus (minus) five centimeters.

To address the issue discussed in the previous section, we emphasize more the choice of covariate sets. However, to show the possibilities of PPE, we want to limit the number of covariates. In covariate selection, it is desirable that the elicitation accounts for infant and adult heights as well as the different growth spurts. For these

Model		**Parameters**							
		b	β_0	β_1					
LINF	$\mathbb{E}[\cdot]$	25.57	50.55	8.84					
$\alpha_{\text{MLE}} = 3$	$\mathbb{SE}[\cdot]$	6.64	2.31	2.18					
		b	h_1	k	t_0				
LOGF	$\mathbb{E}[\cdot]$	31.98	178.11	0.26	3.41				
$\alpha_{\text{MLE}} = 5$	$\mathbb{SE}[\cdot]$	9.56	8.46	0.05	0.11				
		b	h_1	h_{t_*}	s_0	s_1	t_*		
PB	$\mathbb{E}[\cdot]$	43.95	196.92	64.02	0.12	0.09	0.96		
$\alpha_{\text{MLE}} = 20$	$\mathbb{SE}[\cdot]$	13.02	8.95	0.89	0.01	>0.01	0.01		
		b	h_1	C_1	C_2	C_3	D_1	D_2	D_3
JPPS	$\mathbb{E}[\cdot]$	114.89	180.22	0.60	2.61	17.77	3.75	8.82	13.43
$\alpha_{\text{MLE}} = 500$	$\mathbb{SE}[\cdot]$	37.52	5.72	0.02	0.24	3.32	0.08	0.86	0.92
		b	h_1	k	β_0	β_1	c	r	t_*
SSC	$\mathbb{E}[\cdot]$	67.71	1791.66	0.90	827.11	58.26	5.79	10.29	15.67
$\alpha_{\text{MLE}} = 300$	$\mathbb{SE}[\cdot]$	20.87	55.97	0.12	13.26	5.77	0.04	4.66	0.83

Table 6.3: Elicited priors, sampled estimates $\mathbb{E}[\cdot]$ and estimated errors $\mathbb{SE}[\cdot]$ of prior predictive parameters. With each model, the value for the concentration parameter α_{MLE} is reported. The larger α_{MLE} implies better match between the predictive model and the expert expectations.

reasons, we limit the number of covariates used in the elicitation to six, and elicit height probabilities at ages $t = [0, 5, 10, 12, 15, 20]$. The expert probabilities used in elicitation are reported in the Table 6.2.

For this experiment, we used the more accurate gradient-based learning method to ensure the comparability of the results between models. We also optimized the likelihood maximizing parameter α_{MLE} instead of using an approximation. The resulting priors and parameters α_{MLE} are reported in the Table 6.3. As expected, the value of the parameter α_{MLE} is higher for the more sophisticated growth models. Because we used sampling-based computation, the concentration parameters had some variability, thus, we notate the results in the table as approximation. Nevertheless, here the parameter α_{MLE} can be used to distinguish the models without doubt.

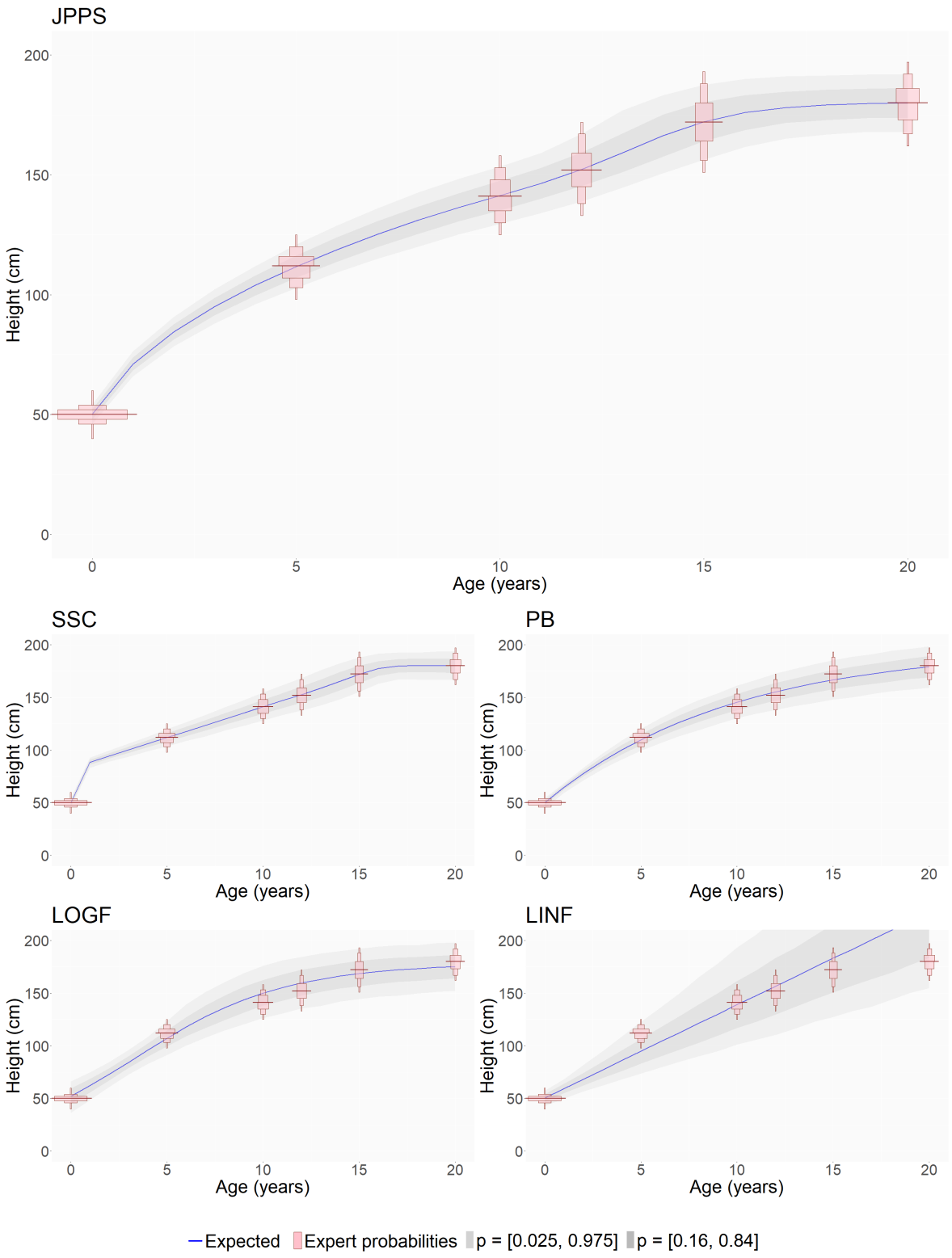


Figure 6.3: Prior predictive growth curves for different models elicited with PPE. For each model, the blue line shows the expected growth curve and the dark and light shaded areas show the prior predictive probabilities at $\mathbb{P}(Y_t) \in [0.16, 0.84]$ and $\mathbb{P}(Y_t) \in [0.025, 0.975]$, respectively. The vertical bar plots show the expert probabilities, and the horizontal line on each plot shows the expert's expectation at the 50th percentile.

Although Ledford and Cole [54] compared the models by fitting them to real data, we have similar finding: the model JPPS ($\alpha_{\text{MLE}} = 500$) has the best matching growth curve w.r.t. expert probabilities, SSC ($\alpha_{\text{MLE}} = 300$) has the second best match and PB ($\alpha_{\text{MLE}} = 20$) has third. The resulting curves for each model are showed in Figure 6.3. There, we can see how the expected curve and shaded areas representing prior predictive probabilities fit the expert probabilities. Note that the infancy growth spurt in the model SSC is visually distinctive, implying that it should have been accounted for better in the choice of covariate sets. Nonetheless, the two most sophisticated models, JPPS and SSC both match the expert data well. In the context of Bayesian workflow, it would be worthwhile to consider either model. Moreover, in some cases the model with more intuitive parameters, like PB or SSC, may be preferred.

This experiment shows that PPE can be used for prior predictive model comparison and analysis given the expert's assessments of the observable data. It should be noted that here the prior informativeness is maximized with respect to expert's prior knowledge. The target here is to find the most appropriate model a priori. However, in a real Bayesian workflow it might be desirable to include uncertainty to the expert knowledge. Recall that in PPE, this is achieved by adjusting the parameter α. Furthermore, when the model is used for posterior analysis after adding observations, the posterior retrodictive checks [10] are still advised.

7. Conclusions

One of the key characteristics of Bayesian machine learning is that the Bayesian models include prior information which is updated with observations. The prior information is described by prior probabilities, and, often it is beneficial that they are based on domain expertise of the modelled issue. We discussed the difficulty of quantifying the expert knowledge and uncertainty in the Chapter 3. In the context of Bayesian models, the elicitation task becomes even more demanding, because the prior knowledge we want to elicit is directly related to the probabilistic structure of a model. Fortunately, we can use a predictive approach to prior elicitation, where the expert is required mainly the domain knowledge and only a small amount of statistical understanding. Here, the expert is asked to give predictive assessments about the observable values, meaning those that are being modelled, instead of structural assessments about the model parameters.

The probabilistic approach to predictive elicitation originally introduced in our recent study [37], and further discussed in this thesis, provides a general, model-independent methodology to perform a predictive prior elicitation. Unlike the predictive approaches in previous literature, the probabilistic predictive elicitation does not require a certain model structure or conjugate prior probability distributions. This becomes apparent in the experiments in Chapter 6, where we used real experts' prior predictive assessments in prior elicitation. Particularly, in Section 6.3 we used the same expert assessments for five different models, and successfully elicited reasonable priors. Furthermore, these experiments show that PPE handles models with multiple parameters well. In addition, we discussed and provided an example in Section 4.3.3 that PPE can be used with multivariate priors.

In Chapter 3, we discussed that while the expert elicitation is worthwhile, it is also prone to errors and uncertainties. The errors often yield from bad practices in elicitation. For example, the expert should be asked for assessments of quantities they are familiar with. Indeed, in the experiment in Section 6.2 we found that the predictive elicitation yielded results that reflected expert knowledge about observable (modelled) quantities much better than the alternative structural method. Moreover, the experts were more confident about their probabilistic assessments regarding observable quan-

tities they were familiar with rather than the model parameters. While this study was limited with few participants, the results were as expected and as such quite indicative. Therefore, among other good elicitation practices, the predictive prior elicitation seems to be useful in reducing elicitation errors in the Bayesian context.

On the other hand, the uncertainties regarding the expert's precision of probabilistic assessments, whether they are due to the lack of expertise, biases or something else, must be considered *after* the probabilities are elicited from the expert. To our knowledge, the proposed methodology is the first to account for the uncertainty relating to the expert's assessments in predictive prior elicitation. This property was discussed with an example in Section 4.3.1. As detailed throughout the Chapter 4, the uncertainty is modelled into PPE through its probabilistic approach, specifically, by the use of the Dirichlet distribution. Specifically, we assume that the probabilities elicited from the expert are a random sample from a certain Dirichlet distribution, that is defined by a concentration parameter and a probability vector. Moreover, this vector is in fact a partitioned prior predictive distribution of the probabilistic model for which we want to find the priors. Ultimately, the goal is to find priors that are the maximum likelihood estimates of the Dirichlet probability density function.

Original in this thesis, we studied further the use cases of PPE. Particularly, we used the properties of Dirichlet distribution in prior predictive model comparison: in the experiment in Section 6.3, we showed how we can compare different models against the expert assessments using the concentration parameter α of the Dirichlet distribution. This is based on the observation discussed in Section 4.1.1, that the approximation of the maximum likelihood estimate of the parameter α is a scaled KL-divergence between the prior predictive probabilities of a model and the expert's elicited probabilities. When we concurrently try to find MLEs for priors and the parameter α, the MLE of the parameter α can be used for model comparison.

Despite the importance, we did not inspect further how the prior distributions of a model with (multiple) parameters are affected by the choice of the Dirichlet concentration parameter α. In future research, it would be interesting to examine for example whether the variance behaves similarly across the prior distributions when the parameter α changes. This is important, because although in the experiments we obtained feasible priors, in Bayesian modelling it is advisable to avoid too informative priors, the issue we discussed in Section 2.2. The challenge emerges if some parameters get very informative i.e. too narrow prior probability distributions and will require excessive amount of observations to update in the posterior analysis. Naturally, even if the elicited priors get very informative, the prior predictive distribution, resulting from elicitation with the discussed methodology, presumably reflects well the experts prior views. Therefore, in the case of very informative priors, essentially the expert is

responsible for providing such assessments.

The experiment results, where we found prior predictive distributions that fit well the experts' expectations, provide interesting opportunities for Bayesian applications. In layered models, such as hierarchical models and Bayesian networks, elicited probabilities have sometimes been used as substitutes for observable parameters, when the observations are scarce. With the use of PPE, we can in theory elicit prior predictive probability distributions for submodels, models which model the observable parameters of the main model, that have limited observations available. Furthermore, we could use the elicited prior predictive model to interpolate and extrapolate expert's knowledge in such layered models. Thus, we could fit in the missing pieces of large Bayesian applications with expert knowledge, and have a sensible priors which the limited amount of observations would update.

In many applications, particularly in those where we need informative priors, it might be desirable to include more than just one expert's knowledge. This task has multiple approaches and requires taking a great care in the elicitation design, as we discussed in Section 3.3. In future research, we could extend the idea of PPE by combining elicited probabilities of multiple experts. Particularly, we could presume that each experts' probabilities are random instances of the same Dirichlet distribution, which is parameterized by the prior predictive model whose priors we want to elicit. Thus, such probabilistic approach could in theory account for differences in experts' knowledge.

In addition to outlining the mathematical machinery of the methodology in Chapter 4, we also covered the computational approaches to implement PPE in practice in Chapter 5. Mostly, we discussed the gradient-based learning methods for finding the maximum likelihood estimates. However, constructing practical applications with such methods required a lot of time, and hence all the examples and experiments excluding the one in Section 6.3 were done using general-purpose optimization tools. Furthermore, despite the theoretical foundation, we could not get the natural gradient work as desired in the stochastic setting. Therefore, we used the *Adam* algorithm to approximate the natural gradient. Moreover, the sampling-based applications caused lengthy runtimes for all optimizers. Fortunately, the tools for more efficient probabilistic modelling and gradient-based learning are continuously evolving.

In this thesis and the research paper [37], we introduced a novel statistical framework that 1) makes prior elicitation independent on the specific structure of the probabilistic model, 2) handles complex models with many parameters and potentially multivariate priors, 3) fully accounts for uncertainty in experts' probabilistic judgements on the data, and 4) provides a formal quality measure indicating if the chosen predictive model is able to reproduce experts' probabilistic judgements. The proposed

methodology adds to the modern Bayesian workflow by providing a simple and precise elicitation methodology, that is model independent, requires the expert only their domain expertise of the observable values, and accounts for the uncertainty of the expert's probabilistic assessments. Furthermore, PPE fits well in the pre-data phase of the Bayesian workflow by providing a principled method for prior checking. Naturally, more work should be done on conceptualizing and assessing the integration of PPE to the modern Bayesian workflow.

Bibliography

[1] A. Akbarov et al. *Probability elicitation: Predictive approach*. PhD thesis, University of Salford, 2009.

[2] S. A. Al-Awadhi and P. H. Garthwaite. An elicitation method for multivariate normal distributions. *Communications in Statistics-Theory and Methods*, 27(5):1123–1142, 1998.

[3] S.-I. Amari. Natural gradient works efficiently in learning. *Neural computation*, 10(2):251–276, 1998.

[4] M. C. Baddeley, A. Curtis, and R. Wood. An introduction to prior information derived from probabilistic judgements: Elicitation of knowledge, cognitive bias and herding. *Geological Society, London, Special Publications*, 239(1):15–27, 2004.

[5] J. Barnard, R. McCulloch, and X.-L. Meng. Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statistica Sinica*, pages 1281–1311, 2000.

[6] M. S. Bartlett. A comment on D. V. Lindley's statistical paradox. *Biometrika*, 44(3/4):533–534, 1957.

[7] E. J. Bedrick, R. Christensen, and W. Johnson. A new perspective on priors for generalized linear models. *Journal of the American Statistical Association*, 91(436):1450–1460, 1996.

[8] E. J. Bedrick, R. Christensen, and W. Johnson. Bayesian binomial regression: Predicting survival at a trauma center. *The American Statistician*, 51(3):211–218, 1997.

[9] J. O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer Science & Business Media, 1985.

[10] M. Betancourt. Towards a principled Bayesian workflow. https://betanalpha.github.io/assets/case_studies/principled_bayesian_workflow.html, 2020. Accessed: 2020-09-15.

[11] P.-C. Bürkner. brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software, Articles*, 80(1):1–28, 2017.

[12] B. Carpenter, A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell. Stan: A probabilistic programming language. *Journal of statistical software*, 76(1), 2017.

[13] G. Casella and R. L. Berger. *Statistical Inference*. Duxbury Press, 2 edition, 2001.

[14] R. T. Clemen, G. W. Fischer, and R. L. Winkler. Assessing dependence: Some experimental results. *Management Science*, 46(8):1100–1115, 2000.

[15] R. T. Clemen and R. L. Winkler. Combining probability distributions from experts in risk analysis. *Risk analysis*, 19(2):187–203, 1999.

[16] R. T. Clemen and R. L. Winkler. Aggregating probability distributions. *Advances in decision analysis: From foundations to applications*, pages 154–176, 2007.

[17] T. Cole. The use and construction of anthropometric growth reference standards. *Nutrition research reviews*, 6(1):19–50, 1993.

[18] N. Dalkey and O. Helmer. An experimental application of the Delphi method to the use of experts. *Management science*, 9(3):458–467, 1963.

[19] A. Daneshkhah and J. Oakley. Eliciting multivariate probability distributions. *Rethinking risk measurement and reporting*, 1, 2010.

[20] E. de Souza da Silva, T. Kuśmierczyk, M. Hartmann, and A. Klami. Prior specification via prior predictive matching: Poisson matrix factorization and beyond. *arXiv*, pages arXiv–1910, 2019.

[21] S. Depaoli. The impact of inaccurate "informative" priors for growth parameters in Bayesian growth mixture modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, 21(2):239–252, 2014.

[22] M. Figurnov, S. Mohamed, and A. Mnih. Implicit reparameterization gradients. In *Advances in Neural Information Processing Systems*, pages 441–452, 2018.

[23] G. Folland. *Real Analysis: Modern Techniques and Their Applications*. Pure and Applied Mathematics: A Wiley Series of Texts, Monographs and Tracts. Wiley, 2013.

[24] H. Fraser, M. Bush, B. Wintle, F. Mody, E. Smith, A. Hanea, E. Gould, V. Hemming, D. Hamilton, L. Rumpff, et al. The repliCATS project - Collaborative Assessments for Trustworthy Science (Predicting reliability through structured expert elicitation with repliCATS). https://replicats.research.unimelb.edu.au/. Accessed: 2021-05-08.

[25] J. Gabry, D. Simpson, A. Vehtari, M. Betancourt, and A. Gelman. Visualization in Bayesian workflow. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 182(2):389–402, 2019.

[26] E. Gaoini, D. Dey, and F. Ruggeri. *Bayesian modeling of flash floods using generalized extreme value distribution with prior elicitation*. University of Connecticut, Department of Statistics, 2009.

[27] P. H. Garthwaite, S. A. Al-Awadhi, F. G. Elfadaly, and D. J. Jenkinson. Prior distribution elicitation for generalized linear and piecewise-linear models. *Journal of Applied Statistics*, 40(1):59–75, 2013.

[28] P. H. Garthwaite and J. M. Dickey. Quantifying expert opinion in linear regression problems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 50(3):462–474, 1988.

[29] P. H. Garthwaite, J. M. Dickey, et al. Elicitation of prior distributions for variable-selection problems in regression. *The Annals of Statistics*, 20(4):1697–1719, 1992.

[30] P. H. Garthwaite, J. B. Kadane, and A. O'Hagan. Statistical methods for eliciting probability distributions. *Journal of the American Statistical Association*, 100(470):680–701, 2005.

[31] A. Gelman. Prior choice recommendations. https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations, 2019. Accessed: 2020-08-26.

[32] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian data analysis*. CRC press, 2014.

[33] A. Gelman and C. Hennig. Beyond subjective and objective in statistics. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 180(4):967–1033, 2017.

[34] A. Gelman, D. Simpson, and M. Betancourt. The prior can often only be understood in the context of the likelihood. *Entropy*, 19(10):555, 2017.

[35] J. P. Gosling. SHELF: The Sheffield Elicitation Framework. In *Elicitation*, pages 61–93. Springer, 2018.

[36] J. P. Gosling, J. E. Oakley, A. O'Hagan, et al. Nonparametric elicitation for heavy-tailed prior distributions. *Bayesian Analysis*, 2(4):693–718, 2007.

[37] M. Hartmann, G. Agiashvili, P.-C. Bürkner, and A. Klami. Flexible prior elicitation via the prior predictive distribution. In *Conference on Uncertainty in Artificial Intelligence (UAI)*. University of Waterloo, 2020.

[38] V. Hemming, M. A. Burgman, A. M. Hanea, M. F. McBride, and B. C. Wintle. A practical guide to structured expert elicitation using the IDEA protocol. *Methods in Ecology and Evolution*, 9(1):169–180, 2018.

[39] J. G. Ibrahim. On properties of predictive priors in linear models. *The American Statistician*, 51(4):333–337, 1997.

[40] J. G. Ibrahim, M.-H. Chen, et al. Power prior distributions for regression models. *Statistical Science*, 15(1):46–60, 2000.

[41] J. G. Ibrahim, M.-H. Chen, Y. Gwon, and F. Chen. The power prior: Theory and applications. *Statistics in medicine*, 34(28):3724–3749, 2015.

[42] A. James, S. L. Choy, and K. Mengersen. Elicitator: An expert elicitation tool for regression in ecology. *Environmental Modelling & Software*, 25(1):129–145, 2010.

[43] I. L. Janis. Groupthink. *Psychology today*, 5(6):43–46, 1971.

[44] P. Jolicoeur, J. Pontier, M.-O. Pernin, and M. Sempé. A lifetime asymptotic growth curve for human height. *Biometrics*, pages 995–1003, 1988.

[45] J. Kadane and L. J. Wolfson. Experiences in elicitation. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 47(1):3–19, 1998.

[46] J. B. Kadane. Predictive and structural methods for eliciting prior distributions. *Bayesian analysis in econometrics and statistics*, pages 89–93, 1980.

[47] J. B. Kadane, N. H. Chan, and L. J. Wolfson. Priors for unit root models. *Journal of Econometrics*, 75(1):99–111, 1996.

[48] J. B. Kadane, J. M. Dickey, R. L. Winkler, W. S. Smith, and S. C. Peters. Interactive elicitation of opinion for a normal linear model. *Journal of the American Statistical Association*, 75(372):845–854, 1980.

[49] R. E. Kass and A. E. Raftery. Bayes factors. *Journal of the american statistical association*, 90(430):773–795, 1995.

[50] R. E. Kass and L. Wasserman. The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, 91(435):1343–1370, 1996.

[51] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[52] D. Kurowicka and R. Cooke. A parameterization of positive definite matrices in terms of partial correlation vines. *Linear Algebra and its Applications*, 372(Supplement C):225–251, 2003.

[53] M. Kynn. The 'heuristics and biases' bias in expert elicitation. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 171(1):239–264, 2008.

[54] A. Ledford and T. Cole. Mathematical models of growth in stature throughout childhood. *Annals of human biology*, 25(2):101–115, 1998.

[55] D. J. Lunn, A. Thomas, N. Best, and D. Spiegelhalter. WinBUGS - a Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and computing*, 10(4):325–337, 2000.

[56] T. P. Minka. Beyond Newton's method, 2000.

[57] T. P. Minka. Estimating a Dirichlet distribution, 2000.

[58] F. A. Moala and A. O'Hagan. Elicitation of multivariate prior distributions: A nonparametric Bayesian approach. *Journal of Statistical Planning and Inference*, 140(7):1635–1655, 2010.

[59] S. Mohamed, M. Rosca, M. Figurnov, and A. Mnih. Monte Carlo gradient estimation in machine learning. *Journal of Machine Learning Research*, 21(132):1–62, 2020.

[60] M. G. Morgan. Use (and abuse) of expert elicitation in support of decision making for public policy. *Proceedings of the National academy of Sciences*, 111(20):7176–7184, 2014.

[61] D. E. Morris, J. E. Oakley, and J. A. Crowe. A web-based tool for eliciting probability distributions from experts. *Environmental Modelling & Software*, 52:1–4, 2014.

[62] J. C. Nash. *optimr: A replacement and extension of the 'optim' function*, 2016.

[63] B. Neuenschwander, M. Branson, and D. J. Spiegelhalter. A note on the power prior. *Statistics in medicine*, 28(28):3562–3566, 2009.

[64] M. Nevalainen, I. Helle, and J. Vanhatalo. Estimating the acute impacts of Arctic marine oil spills using expert elicitation. *Marine pollution bulletin*, 131:782–792, 2018.

[65] J. Oakley. *SHELF: Tools to support the Sheffield Elicitation Framework*, 2019. R package version 1.6.0.

[66] J. E. Oakley and A. O'Hagan. Uncertainty in prior elicitations: A nonparametric approach. *Biometrika*, 94(2):427–441, 2007.

[67] A. O'Hagan. Eliciting expert beliefs in substantial practical applications. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 47(1):21–35, 1998.

[68] A. O'Hagan. Expert knowledge elicitation: Subjective but scientific. *The American Statistician*, 73(sup1):69–81, 2019.

[69] A. O'Hagan, C. E. Buck, A. Daneshkhah, J. R. Eiser, P. H. Garthwaite, D. J. Jenkinson, J. E. Oakley, and T. Rakow. *Uncertain judgements: Eliciting experts' probabilities*. John Wiley & Sons, 2006.

[70] A. O'Hagan and J. E. Oakley. Probability is perfect, but we can't elicit it perfectly. *Reliability Engineering & System Safety*, 85(1-3):239–248, 2004.

[71] A. Parducci. Range-frequency compromise in judgment. *Psychological Monographs: General and Applied*, 77(2):1, 1963.

[72] D. F. Percy. Bayesian enhanced strategic decision making for reliability. *European Journal of Operational Research*, 139(1):133–145, 2002.

[73] D. F. Percy. Subjective reliability analysis using predictive elicitation. In *Mathematical and Statistical Methods in reliability*, pages 57–72. World Scientific, 2003.

[74] D. F. Percy. Subjective priors for maintenance models. *Journal of quality in maintenance engineering*, 2004.

[75] J. Piironen and A. Vehtari. Comparison of Bayesian predictive methods for model selection. *Statistics and Computing*, 27(3):711–735, 2017.

[76] M. Plummer et al. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd international workshop on distributed statistical computing*, volume 124, pages 1–10. Vienna, Austria, 2003.

[77] M. Preece and M. Baines. A new family of mathematical models describing the human growth curve. *Annals of human biology*, 5(1):1–24, 1978.

[78] A. Sarma and M. Kay. Prior setting in practice: Strategies and rationales used in choosing prior distributions for Bayesian analysis. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2020.

[79] D. J. Schad, M. Betancourt, and S. Vasishth. Toward a principled Bayesian workflow in cognitive science. *Psychological methods*, 2020.

[80] C. S. Spetzler and C.-A. S. Staël von Holstein. Probability encoding in decision analysis. *Management science*, 22(3):340–358, 1975.

[81] L. Suantak, F. Bolger, and W. R. Ferrell. The hard–easy effect in subjective probability calibration. *Organizational Behavior and Human Decision Processes*, 67(2):201–221, 1996.

[82] A. Tversky and D. Kahneman. Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157):1124–1131, 1974.

[83] A. Tversky and D. Kahneman. The framing of decisions and the psychology of choice. *Science*, 211(4481):453–458, 1981.

[84] University of Eastern Finland and Kuopio University Hospital. Growth curves / Boys / 0-2 years and 1-20 years (*Kasvukäyrät / Pojat / 0-2 vuotiaat ja 1-20 vuotiaat*). http://kasvukayrat.fi/paperikayrat/, 2018. Accessed: 2021-03-24.

[85] A. Vehtari, J. Ojanen, et al. A survey of Bayesian predictive methods for model assessment, selection and comparison. *Statistics Surveys*, 6:142–228, 2012.

[86] R. L. Winkler. The assessment of prior distributions in Bayesian analysis. *Journal of the American Statistical association*, 62(319):776–800, 1967.

[87] R. L. Winkler. Prior information, predictive distributions, and Bayesian model-building. *Bayesian analysis in econometrics and statistics: Essays in honor of Harold Jeffreys*, pages 95–109, 1980.

[88] J. Xuan, J. Lu, and G. Zhang. A survey on Bayesian nonparametric learning. *ACM Computing Surveys (CSUR)*, 52(1):1–36, 2019.

Appendix A. Technical details of the implementation of PPE

The implementation of the probabilistic predictive elicitation consists of three key components: elicitation of expert's predictive assessments, construction of a probabilistic model, and optimization of hyperparameters that maximize the Dirichlet log-likelihood. In our implementation, we use the `R` programming language, and, we rely on existing programming libraries to manage the expert elicitation and to operate the probabilistic model. Particularly, we use the `SHELF` [65] package for the expert elicitation, and the `brms` [11] package to construct probabilistic models. In the optimization task we explore different options for implementation. The general layout of our approach is illustrated in the Figure A.1.

For probabilistic modelling, we use the `brms` [11] package, which fits Bayesian generalized (non-)linear multivariate multilevel models using the probabilistic programming language Stan. The models are fitted using Markov chain Monte Carlo (MCMC) methods, meaning that the probabilistic models are sampled and stochastic. In our implementation, we focus solely on Bayesian regression models which are structured as

$$\begin{aligned} Y_x|\,\boldsymbol{\theta}, \boldsymbol{b} &\sim \mathcal{P}_Y(f(\boldsymbol{x};\boldsymbol{\theta}), \boldsymbol{b}) \\ b_i &\sim \mathcal{P}_i(\boldsymbol{\gamma}_i) \\ \theta_j &\sim \mathcal{P}_j(\boldsymbol{\lambda}_j), \end{aligned}$$

where Y_x is the observed dependent variable given the covariate x which follows a probability distribution $\mathcal{P}_{\mathcal{Y}}(f(\boldsymbol{x};\boldsymbol{\theta}), \boldsymbol{b})$. There, $f(\boldsymbol{x};\boldsymbol{\theta})$ is the observational model that is parameterized by the parameter vector $\boldsymbol{\theta}$ and takes the covariate vector $\boldsymbol{x}$ as an input. The vector $\boldsymbol{b}$ are possible additional parameters of the probability distribution $\mathcal{P}_{\mathcal{Y}}$. Each parameter in vectors $\boldsymbol{b}$ and $\boldsymbol{\theta}$ follow parametric probability distributions, and vectors $\boldsymbol{\gamma}$ and $\boldsymbol{\lambda}$ are their hyperparameters. Ultimately, the hyperparameters define prior distributions of model parameters. Thus, the goal of our application is to find hyperparameters which define a prior predictive model which corresponds to expert probabilities. An example of how we set the probabilistic model with `brms` is shown in the code snippet in Figure A.2.

For elicited expert probabilities we use the same data structure as in `SHELF` [65]. Particularly, we use a light modification of the *Multiple experts* graphical user interface, where instead of multiple experts we ask one expert to provide probabilities for multiple covariates or covariate sets. The interface allows easy documentation of the elicitation results, however, it does not restrict how the elicitation is conducted in practice. The Figure A.3 illustrates how the expert probabilities of Section 6.3 are uploaded in the

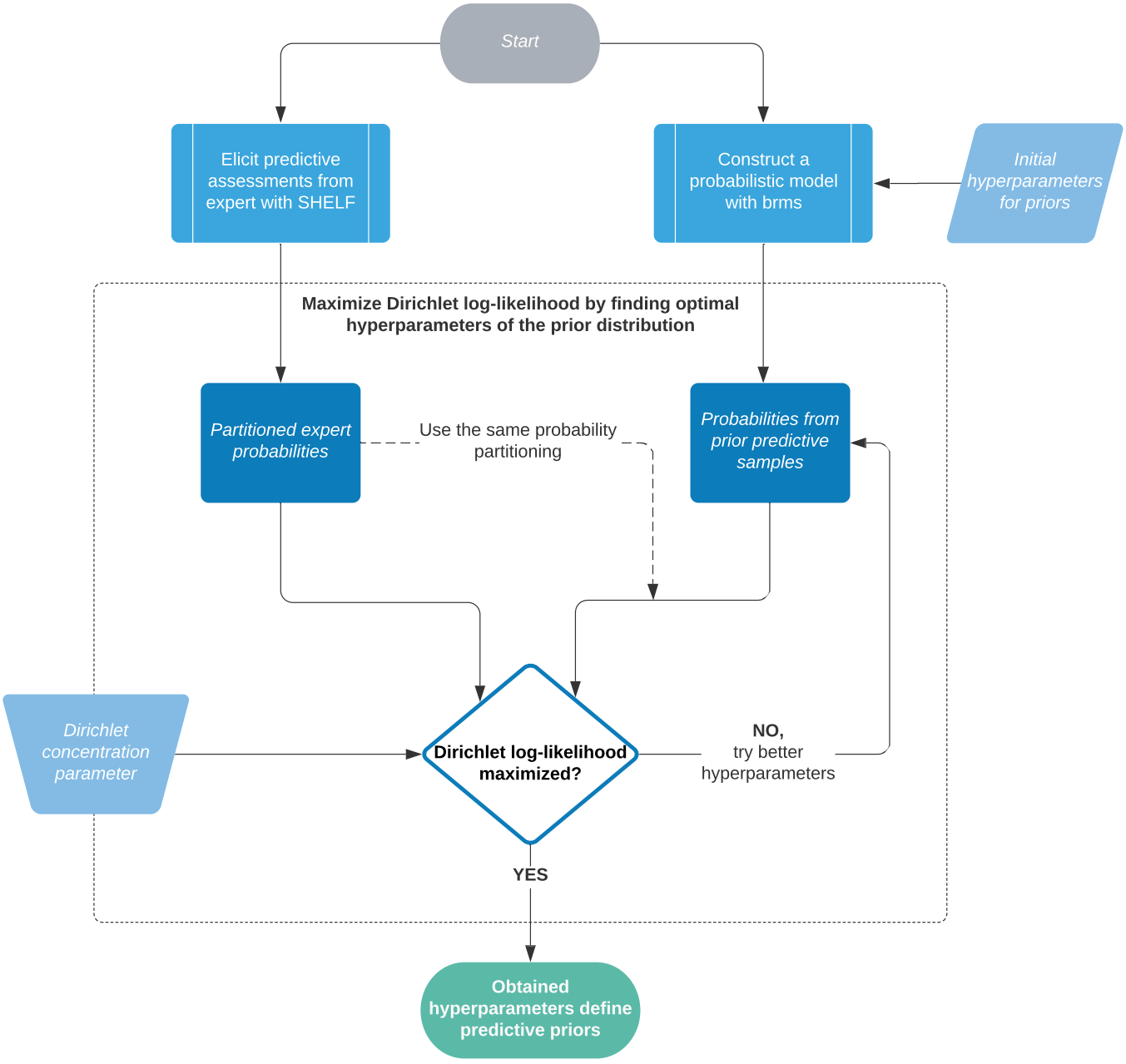


Figure A.1: The implementation of PPE in `R` language summarized to a flowchart. We begin with eliciting expert probabilities in a (modified) `SHELF` interface, and constructing a probabilistic model with the `brms` package. The optimization phase, illustrated as the box with dotted outline, finds the hyperparameters that maximize the Dirichlet log-likelihood discussed in Chapter 4. Because the expert probabilities are partitioned, we also partition the sampled prior predictive distribution accordingly. The Dirichlet concentration parameter can be set to a constant or it can be optimized. If it is optimized, the optimization can be done concurrently or alternately with the optimization of model's hyperparameters.

```
# 1) Define the regression formula
bform <- bf(h ~ h0 + r * t,
            h0 + r ~ 1,
            nl = TRUE)

# 2) Define covariates used in the elicitation, and arbitrary response variables
datPPE <- data.frame(h=c(50, 110, 140, 150, 160, 180),
                     t=c(0, 5, 10, 12, 15, 20))

# 3) Define prior distributions, where hyperparameters are named variables
(priors <-
    prior(gamma(gamma1, gamma2), class = "shape") +
    prior(lognormal(lna_h0, lnb_h0), nlpar = "h0", lb=0, class="b") +
    prior(lognormal(lna_r, lnb_r), nlpar = "r", lb=0, class="b")
)

# 4) Define stanvar objects for named hyperparameters with initial values
stanvars <- stanvar(8, name = "gamma1") +
  stanvar(1, name = "gamma2") +
  stanvar(4, name = "lna_h0") +
  stanvar(0.05, name = "lnb_h0") +
  stanvar(2, name = "lna_r") +
  stanvar(0.3, name = "lnb_r")

# 5) Sample from brms model only the prior predictive model.
#    The probability distribution P_Y and link function are defined here
model <- brm(bform,
             data = datPPE,
             prior = priors,
             family = weibull(link = "identity"),
             seed = 42,
             sample_prior = "only",
             stanvars = stanvars,
             iter=2000)
```

Figure A.2: Example of creating a probabilistic model in `brms`. Here, we define the model LINF from the Section 6.3. In addition to the covariates we use also in the expert elicitation, `brms` requires passing response variables. However, the defined response variables do not affect prior predictive sampling. Using *stanvar* objects allows us not to recompile the `Stan` model each time we change the hyperparameters and generate new samples. To keep the programming structure simple and flexible, we set the parameter *nl* (non-linear) always as TRUE.

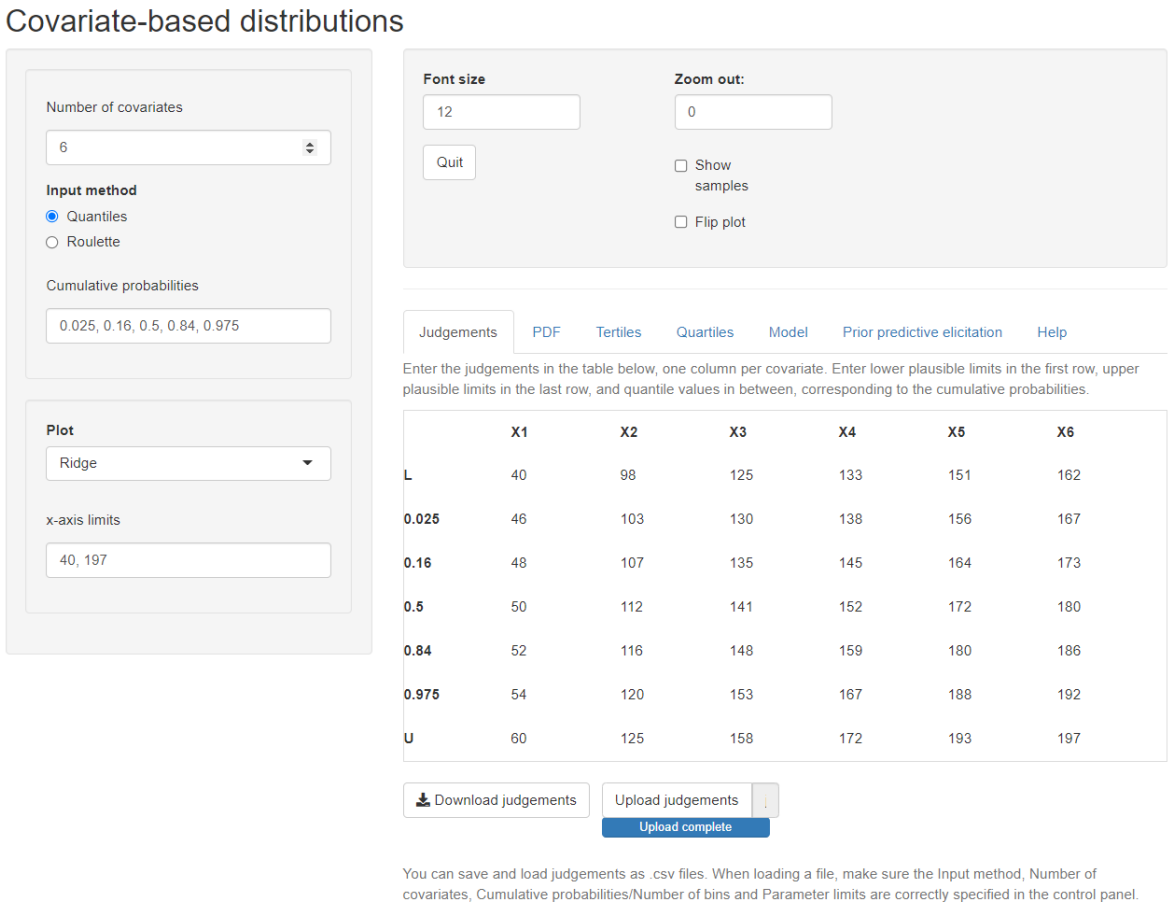
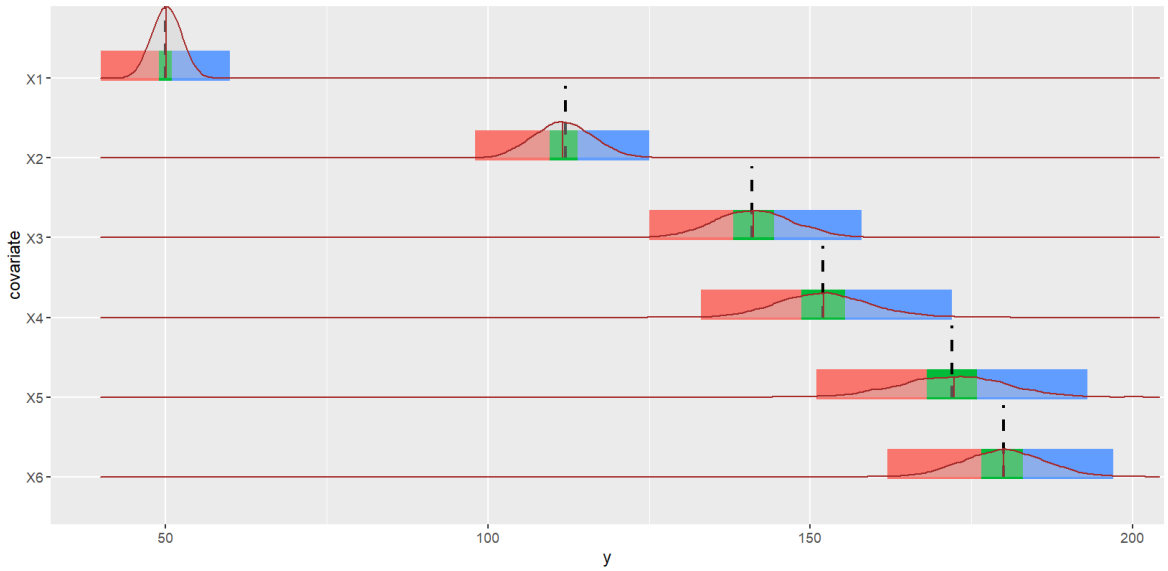


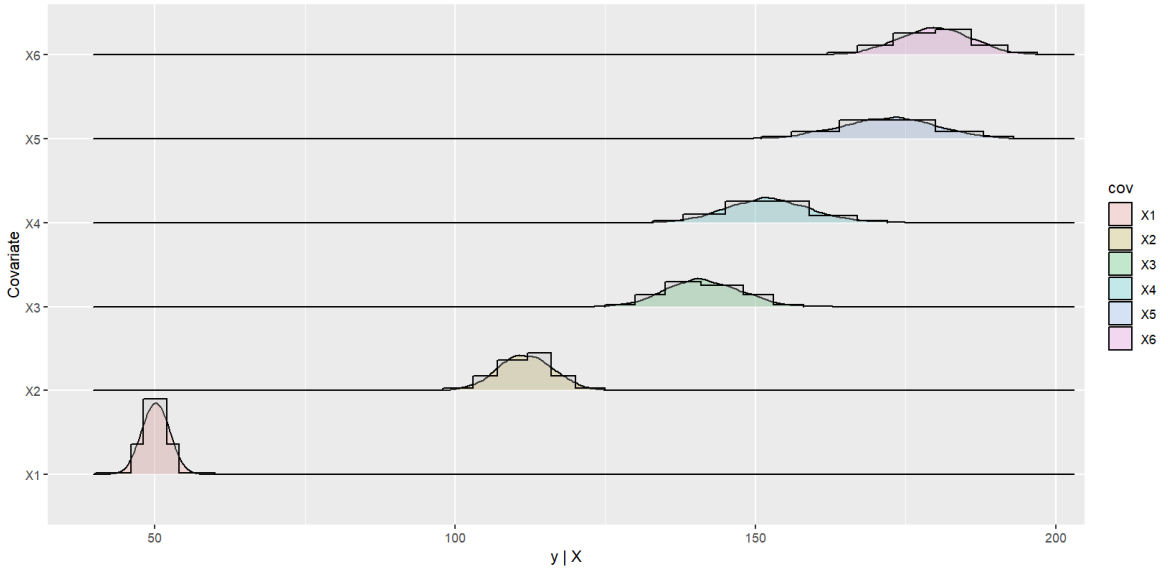
Figure A.3: The graphical user interface of the *Multiple experts* method in `SHELF` modified for multiple covariates in probabilistic predictive elicitation.

interface. The probabilistic predictive elicitation can be conducted through the interface, but we achieved better computational stability when running the optimization tools separately from the graphical interface. Nevertheless, the interface can be used for visual comparison of the model probabilities and expert probabilities as the plots in Figure A.4 show.

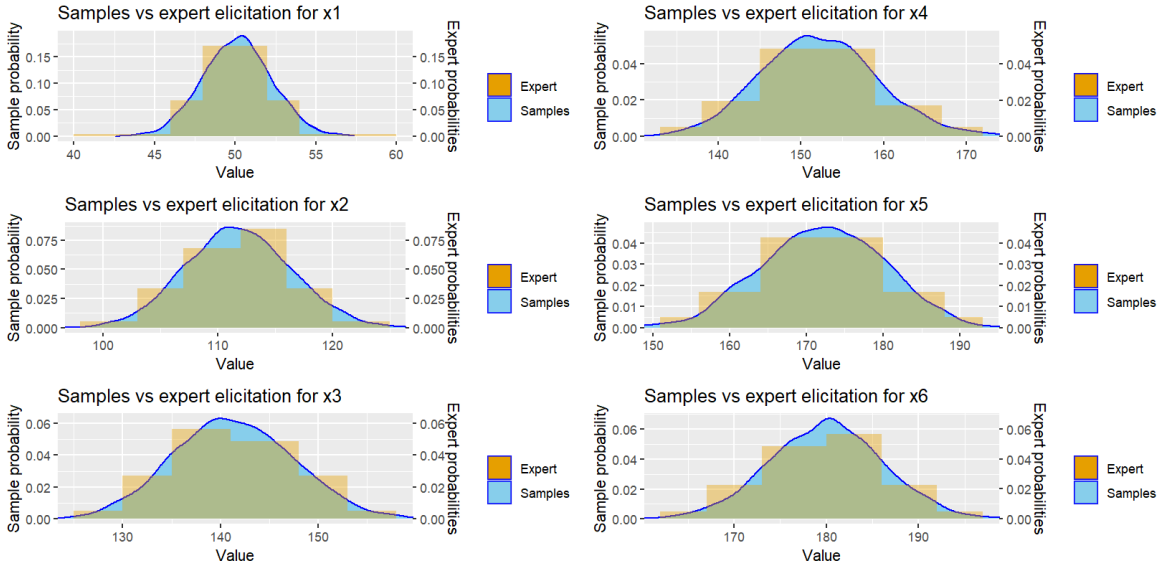
Once the expert probabilities are elicited and the probabilistic model is built with initial hyperparameters, we move to the optimization phase which is illustrated inside the dotted box in the flowchart in Figure A.1. The optimization tools in `R` often have a similar interface, which take as an input the target function and the parameters that are being optimized. In our implementation, the target function consists of three parts. First, the prior predictive distribution is sampled given the hyperparameters. Second, the sampled data is partitioned into a discrete set of probabilities according to the partitioning of the expert probabilities. Finally, the sampled probabilities and expert probabilities are used to calculate the log-likelihood of the Dirichlet density function shown in Chapter 4, Equation (4.2). For computational purposes, the target function



(a) Tertiles plot from `SHELF` with the prior predictive density added.



(b) Ridgeline plot of expert probabilities and prior predictive densities at each covariate.



(c) Density plots of expert probabilities and prior predictive densities at each covariate.

Figure A.4: Example plots included in the graphical interface for the comparison of the expert probabilities (bars) and prior sampled predictive probability densities at each covariate. The example shows the results of the model JPPS in Section 6.3.

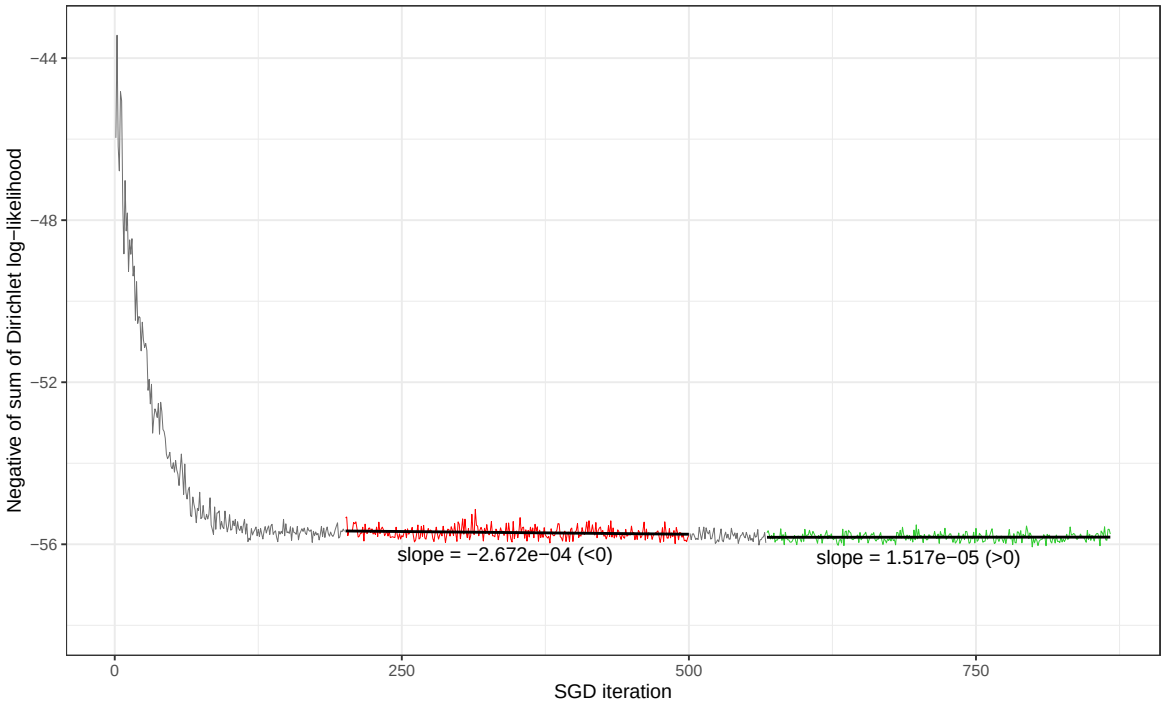


Figure A.5: The stopping rule of the implemented stochastic gradient descent illustrated. The y-axis shows the result of a target function being minimized and the x-axis shows the number of iterations. After the first 500 iterations, we compute the slope of a linear model fitted with last 300 results. The optimization stops when the slope is positive. The jagged line shows the result of a target function, and the straight lines show the fitted linear model when the slope is computed for the first time (red samples) and the last time (green samples), on which the stopping rule activates.

is often minimized, and since we have covariate-based models, ultimately the target function returns the negative of the sum of Dirichlet log-likelihoods. The optimized parameters are naturally the hyperparameters, and, optionally the concentration parameter of the Dirichlet distribution. Furthermore, the concentration parameter can also be set as a constant or optimized alternately with the hyperparameters.

The so-called blackbox optimization methods we used successfully include Nelder-Mead, NEWUOA and Bayesian optimization. However, they could not reach a very high accuracy with a stochastic model such as ours. The accuracy becomes particularly important in the model comparison, for example in the one conducted in the Section 6.3. Therefore, for that experiment we built a stochastic gradient descent (SGD) optimizer which is described at a theoretical level in Chapter 5. Although instead of using natural gradient, we used the Adam algorithm [51] that simulates the natural gradient in a stochastic environment. The computation of gradients was customized for our models, so that the gradient of the observational model was calculated using a symbolic differentiation of the package `Ryacas`, and the gradients of probability distributions were computed using the reparameterization trick described in Section 5.1.2.

When sampling the prior predictive distribution with `brms`, we use the default settings: four chains and 2,000 iterations per chain of which half are warmup (burnin) samples. In the stochastic gradient descent, we use a batch of 150 random sampled (of the total 4,000 after removing warmup) to compute the gradient. We vary the learning rate (step size) of the gradient descent between 0.001 and 0.1 depending on the performance, and for the Adam algorithm's learning parameters we use the default values recommended by the authors [51]. The SGD runs for at least 500 iterations, and stops when the slope of a fitted linear model computed with a rolling 300 target function results becomes positive, or after 10,000 iterations. The Figure A.5 illustrates the stopping rule used. When using a gradient-based optimization to maximize the Dirichlet likelihood, the Dirichlet concentration parameter must be set constant or optimized alternately with other parameters, because the gradients of the concentration parameter and other parameters depend on each other.

Because we build probabilistic models with `brms` which uses MCMC sampling, the optimization methods are computationally expensive. First, at each iteration of any optimizer the prior predictive distribution must be sampled. Second, the samples are stochastic which may confuse some optimization tools. The stochastic gradient descent handles stochasticity well, but our implementation is likely not optimal. While this implementation works as a proof of concept, it would be advisable to build probabilistic models with tools that support stochastic gradient descent or use a wrapper such as `Stan2tfp`, that allows compiling a `Stan` language model into the `TensorFlow`, a software which supports differentiable programming. The current implementation would not be ideal for interactive elicitation, where the expert can see the elicited prior predictive distribution straight away, and possibly react to it by adjusting their assessment.

Appendix B. Height growth elicitation results

In Section 6.2, we described the real world experiment using probabilistic predictive elicitation as done in the original study [37]. The main text provided the elicitation results for one example participant. For the rest, the results are provided here tables B.1 to B.4 similar to the appendix of the original study. For all participants, the predictive prior elicitation produced higher value for the concentration parameter α, implying that the predictive results diverged less from the predictive expert probabilities. Moreover, the predictively elicited priors are a closer match to the reference values of the data-dependent values in the Preece and Baines' study [77].

Table B.1: Participant 2

		Predictive		**Parametric**	
Parameter	Reference	$\mathbb{E}[\cdot]$	$\mathbb{V}(\cdot)$	$\mathbb{E}[\cdot]$	$\mathbb{V}(\cdot)$
h_1	174.6	191.74	4.32	172.7	101.6
h_{t_*}	162.9	153.73	1.6	129.1	31.0
s_0	0.1	0.04	< 0.01	0.51	< 0.04
s_1	1.2	2	4.3	0.5	< 0.04
t_*	14.6	15.9	0.7	12.9	0.5
b		61.4	111.4	3.1	2.6
α	–	14.0	–	1.3	–

Table B.2: Participant 3

		Predictive		**Parametric**	
Parameter	Reference	$\mathbb{E}[\cdot]$	$\mathbb{V}(\cdot)$	$\mathbb{E}[\cdot]$	$\mathbb{V}(\cdot)$
h_1	174.6	177.14	3.68	174.6	146.3
h_{t_*}	163.0	148.8	1.86	78.5	37.2
s_0	0.1	0.07	< 0.001	0.2	0.004
s_1	1.2	4.54	37.83	0.9	0.004
t_*	14.6	11.31	0.21	6.9	2.9
b	–	18.4	12.5	25.8	74.1
α	–	9.5	–	1.5	–

Table B.3: Participant 4

		Predictive		**Parametric**	
Parameter	Reference	$\mathbb{E}[\cdot]$	$\mathbb{V}(\cdot)$	$\mathbb{E}[\cdot]$	$\mathbb{V}(\cdot)$
h_1	174.6	174.5	< 0.01	50.5	64.5
h_{t_*}	162.9	162.8	0.02	129.1	31.0
s_0	0.1	0.1	< 0.01	5.1	2.7
s_1	1.2	1.6	1.7	5.1	2.7
t_*	14.60	14.7	0.9	12.9	0.6
b	–	14.5	14.3	1	< 0.02
α	–	17.1	–	1.2	–

Table B.4: Participant 5

		Predictive		**Parametric**	
Parameter	Reference	$\mathbb{E}[\cdot]$	$\mathbb{V}(\cdot)$	$\mathbb{E}[\cdot]$	$\mathbb{V}(\cdot)$
h_1	174.6	174.4	0.91	159.66	155.96
h_{t_*}	162.9	162.6	0.85	121.75	57.27
s_0	0.1	0.1	< 0.01	3.3	3.3
s_1	1.2	3.4	< 0.01	3.3	3.3
t_*	14.6	14.6	0.02	11.7	5.36
b	–	17.8	17.8	9.5	8.3
α	–	7.7	–	1.5	–